Analyse et caractérisation temps réel de vidéos chirurgicales. Application à la chirurgie de la cataracte

Katia Charriere

THÈSE / Télécom Bretagne
sous le sceau de l'Université européenne de Bretagne

pour obtenir le grade de Docteur de Télécom Bretagne
En accréditation conjointe avec l'Ecole Doctorale Sicma
Mention : Sciences et Technologies de l'Information et de la Communication

présentée par
Charrière Katia

préparée dans le département Image et traitement de l'information
Laboratoire Latim

Analyse et caractérisation temps réel de vidéos chirurgicales. Application à la chirurgie de la cataracte

Thèse soutenue le 23 novembre 2015
Devant le jury composé de :

Pascal Haigron
Professeur, Université de Rennes I / président

Michel de Mathelin
Professeur, Directeur d'iCube, Télécom-Physique Strasbourg / rapporteur

Pierre Jannin
Directeur de Recherche, Medicis, Faculté de médecine - Rennes / rapporteur

Béatrice Cochener
Professeur des Universités - Praticien Hospitalier, Chef du service d'Ophtalmologie, CHRU de Brest / examinateur

Guy Cazuguel
Directeur d'études, Télécom Bretagne / examinateur

Gouenou Coatrieux
Professeur, Télécom Bretagne / directeur de thèse

Mathieu Lamard
Ingénieur de recherche, Université de Bretagne Occidentale / invité

Gwénolé Quellec
Chargé de recherche Inserm - Brest / invité

N° d'ordre : 2015telb0380

Sous le sceau de l'Université européenne de Bretagne

Télécom Bretagne

En accréditation conjointe avec l'Ecole Doctorale Sicma

Analyse et caractérisation temps réel de vidéos chirurgicales. Application à la chirurgie de la cataracte

---

Thèse de Doctorat

Mention : Maths – STIC (sciences et technologies de l'information et de la communication)

Présentée par **Katia Charrière**

Département : Image et traitement de l'information

Laboratoire : LaTIM – INSERM UMR 1101

Directeurs de thèse : M. Gouenou Coatrieux et Mme Béatrice Cochener

Soutenue le 23 Novembre 2015

Jury :

Rapporteurs :	M. **Michel de Mathelin**, Professeur, Directeur d'iCube, Télécom-Physique Strasbourg
	M. **Pierre Jannin**, Directeur de Recherche INSERM, Faculté de médecine, Rennes
Examinateurs :	Mme **Béatrice Cochener**, Professeur des Universités - Praticien Hospitalier, Chef du service d'Ophtalmologie, CHRU de Brest
	M. **Gouenou Coatrieux**, Professeur, Télécom Bretagne
	M. **Pascal Haigron**, Professeur, Université de Rennes I
	M. **Guy Cazuguel**, Directeur d'Etudes, Télécom Bretagne
Invités :	M. **Mathieu Lamard**, Ingénieur de Recherche, Université de Bretagne Occidentale
	M. **Gwénolé Quellec**, Chargé de Recherche INSERM, LaTIM, Brest

Résumé

Le domaine médical génère des données en grande quantité dont l'exploitation peut être un véritable atout pour l'aide à la pratique médicale. Le LaTIM a développé de solides compétences dans l'analyse automatique de ces données, notamment pour l'aide au diagnostic. Dans le cadre de cette thèse nous proposons de nous appuyer sur des vidéos chirurgicales préalablement archivées et interprétées pour apporter une aide opératoire en temps réel au chirurgien. Pour que cette aide soit pertinente, il est tout d'abord nécessaire de reconnaître, à chaque instant, le geste pratiqué par le chirurgien. Ce point est essentiel et fait l'objet de cette thèse. Les méthodes développées sont évaluées sur une base de données de vidéos de chirurgies de la cataracte, collectées grâce à une forte collaboration avec le service d'ophtalmologie du CHRU de Brest.

Nous nous ramenons tout d'abord à un problème de catégorisation de séquences vidéo, où chaque séquence représente une tâche chirurgicale. La catégorisation de ces tâches s'appuie sur la recherche de cas similaires dans la base de données. Pour cela, nous utilisons le mouvement extrait de la vidéo pour construire des signatures visuelles. Nous utilisons une mesure de similitude, alternative à l'algorithme DTW, proposée par Piciarelli et al. [1] pour la surveillance du trafic routier. Les performances de catégorisation, avec une aire moyenne sous la courbe ROC $A_z = 0{,}794$ sont satisfaisantes au vu de la rapidité des algorithmes.

Nous cherchons ensuite à analyser en temps réel la chirurgie en cours d'acquisition et à reconnaître à chaque instant le geste chirurgical réalisé. La vidéo requête est alors considérée comme une succession de sous-séquences pouvant se chevaucher. Nous apportons en plus du contenu visuel de la sous séquence en cours, une information contextuelle en nous appuyant sur un modèle statistique du processus chirurgical de la chirurgie pour labelliser la séquence. Un premier modèle sous forme d'arbre évalué n'a pas donné de résultats satisfaisants. Un second modèle de la chirurgie évalué, à plusieurs niveaux de description, tire avantage des relations de cause à effet qui existent entre les niveaux de description ainsi que des connaissances du déroulement temporel de la chirurgie pour affiner la reconnaissance. Le modèle peut prendre en entrée différents types de preuves. Des résultats encourageants ont été obtenus en utilisant comme preuve le résultat d'une recherche de cas similaires ($A_z = 0{,}759$). La pertinence du modèle a en outre été validée en utilisant comme preuve une information fournie par les chirurgiens : les instruments utilisés dans la sous séquence requête ($A_z = 0{,}983$).

Des résultats encourageants sont obtenus pour la reconnaissance automatique du geste chirurgical et le modèle statistique multi-échelles permet une analyse fine et complète de la chirurgie. Il a été validé et l'information sur la présence et le type des instruments utilisés, fournie actuellement par les chirurgiens, pourra être obtenue par la suite grâce à une détection automatique. L'approche proposée est très générale et devrait permettre d'alerter le chirurgien sur les déroulements opératoires à risques, et lui fournir des recommandations en temps réel sur des conduites à tenir reconnues. Les méthodes développées permettront également d'indexer automatiquement des vidéos chirurgicales.

Mots clés : Recherche de vidéos par le contenu, modélisation multi-échelles, analyse du processus chirurgical, modèles de Markov, champs markoviens conditionnels, réseau bayésien

Abstract

Huge amounts of medical data are recorded every day. Those data could be really helpful for medical practice. The LaTIM has been developing solid skills about the analysis of those data for decision support. In this PhD thesis, we propose to reuse annotated surgical videos previously recorded and stored in a dataset, for computer-aided surgery. That implies to first know at each instant of the surgery, which surgical gesture is being performed, to be able to provide relevant information. This challenging task is the aim of this thesis. The methods developed during this thesis have been evaluated in a video dataset of cataract surgeries, collected at Brest University Hospital.

We first work on real-time recognition of cataract surgery tasks: the goal is to retrieve, from a database, surgery videos that were recorded during the same surgery task. A content based video retrieval (CBVR) method is used to categorize the query video. The proposed system relies on motion feature for video characterization and a similarity measure adapted from Piciarelli's video surveillance system is evaluated. This similarity measurement is an alternative to the DTW algorithm. With a mean area under the ROC curve $A_z = 0.794$, retrieval performances are satisfactory in regard to the low computation times.

We then propose an automated analysis of cataract surgeries, in real time, during the video recording. We try to recognize, at each instant of the surgery, which surgical gesture is being performed. The video is sequenced into fixed sized overlapping sub-sequences. In addition to the use of visual content of each sub-sequence, we also use a statistical model of surgical process to bring contextual information. A tree model of the cataract surgery has been evaluated, with not satisfactory results. A second one was used, based on a multilevel description of the surgery. This model uses the conditional relationships between each level of description and temporal relationship to refine the labeling of each sequence. Observations taken as input of the model derive either from the presence of surgical instruments or from the analysis of motion in videos through the CBVR paradigm. Promising results were obtained using motion analysis ($A_z = 0.759$) and the information about the presence of surgical instruments, manually provided by the surgeons, validated the model with a mean area under the ROC curve $A_z = 0.983$.

Promising results were obtained for automated analysis of cataract surgeries and surgical gesture recognition. The statistical model allows an analysis which is both fine-tuned and comprehensive. The presence of surgical instruments, which provides good results, are currently manually given by surgeons, but this step could be automated by tracking and recognition methods. The general approach proposed in this thesis could be easily used for computer aided surgery, by providing recommendations or video sequence examples. The method could also be used for video annotation for database indexing.

Keywords: content based video retrieval, multilevel statistical model, surgical process model, Markov models, conditional random fields and Bayesian networks

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

Remerciements

Remerciements

En premier lieu, je tiens à remercier **le LaTIM**, dirigé par **M. Eric Stindel**, de m'avoir accueillie pour cette thèse.

Cette thèse est le résultat d'un vrai travail d'équipe et je voudrais tout d'abord remercier mes directeurs de thèse. Je remercie ainsi **Gouenou Coatrieux**, de m'avoir guidée dans cette thèse avec des mots justes et pertinents, et **Béatrice Cochener** de m'avoir offert l'opportunité de travailler avec un grand nombre de données chirurgicales et de nous avoir fait part de toute son expertise.

Je tiens à exprimer ma profonde gratitude à mes encadrants. Tout d'abord à **Guy Cazuguel**, merci de m'avoir donné l'opportunité de réaliser cette thèse et de m'avoir guidée avec tant de justesse et de gentillesse. Je remercie également tout particulièrement **Gwénolé Quellec**, pour son aide précieuse, sa disponibilité et sa patience. Ses conseils ont été essentiels pour la réalisation de ce travail de thèse. Ses compétences et sa rigueur scientifique m'ont beaucoup appris. Enfin, je remercie **Mathieu Lamard**, pour avoir été de justes conseils tout au long de ces trois années. A vous trois, merci pour votre confiance et votre gentillesse, c'était un réel plaisir et une chance d'avoir travaillé avec vous. Enfin, je remercie **David Martiano** pour cette collaboration très enrichissante.

Merci à Brest métropole et à l'Institut Télécom pour le financement de cette thèse.

Je remercie également **M. Pierre Jannin**, **M. Michel de Mathelin** et **M. Pascal Haigron** d'avoir accepté d'examiner mon travail.

Merci à tous les membres du LaTIM et toutes les personnes que j'ai pu rencontrer pendant cette thèse.

Remerciements

Sommaire

RESUME **3**

ABSTRACT **4**

REMERCIEMENTS **6**

SOMMAIRE **8**

TABLE DES FIGURES **12**

TABLE DES TABLEAUX **16**

GLOSSAIRE **18**

INTRODUCTION **20**

CHAPITRE I. CONTEXTE **22**

I.1 Réutilisation des archives médicales numériques **23**
I.1.1 Les archives médicales numériques 23
I.1.2 La fouille de données 24
I.1.2.1 Le raisonnement à partir de cas 25
I.1.2.2 La Recherche d'images par le contenu (CBIR) 26
I.1.2.3 La recherche de vidéos par le contenu (CBVR) 27
I.1.3 Positionnement du travail de thèse dans les recherches du LaTIM 28
I.1.3.1 La Recherche d'images par le contenu (CBIR) 29
I.1.3.2 L'analyse automatique de vidéos chirurgicales 30
I.1.4 Conclusion 30

I.2 Aide à la chirurgie en temps réel **31**
I.2.1 Travaux principaux en analyse automatique de vidéos 31
I.2.1.1 Dans le domaine médical 31
I.2.1.1.1 Résumé automatique d'examens 31
I.2.1.1.2 Annotation automatique de vidéos 32
I.2.1.1.3 Recherche de cas similaires 33
I.2.1.1.4 Evaluation des compétences des chirurgiens 33
I.2.1.1.5 L'analyse de scènes opératoires 34
I.2.1.1.6 Conclusion 34
I.2.1.2 Analyse de séquences vidéos dans d'autres domaines 36
I.2.1.3 Conclusion 37
I.2.2 Les travaux du LaTIM dans l'analyse automatique de vidéos chirurgicales 37
I.2.2.1 Reconnaissance automatique de tâches chirurgicales 37

I.2.2.2 Séquençage automatique de vidéos de chirurgies 40
I.2.3 L'analyse automatique de vidéos en temps réel 41
I.2.3.1 Algorithmes rapides 41
I.2.3.2 Algorithmes d'analyse « en direct » 41
I.2.3.3 Conclusion 42
I.2.4 Scénario envisagé 42
I.2.4.1 Reconnaissance du geste chirurgical 42
I.2.4.2 Détection d'événements anormaux et génération d'alertes 44

I.3 Discussion - Conclusion 45

CHAPITRE II. BASES DE DONNEES 46

II.1 Les bases de données et leurs description dans la littérature 47
II.1.1 Granularité 47
II.1.2 Les différentes descriptions des bases de données de vidéos médicales 48
II.1.2.1 Travaux de l'équipe VisAGeS 49
II.1.2.2 Autres travaux 49

II.2 La base de données du LaTIM 53
II.2.1 Application clinique : la chirurgie de la cataracte 53
II.2.2 Les données 54
II.2.2.1 Description de la chirurgie 54
II.2.2.2 Les bases de données annotées 56

II.3 Nouvelle description multi-échelles de la chirurgie 57
II.3.1 Phases 57
II.3.2 Etapes 58
II.3.3 Activités 59

II.4 Diagrammes de transition obtenus 60
II.4.1 Phases 60
II.4.2 Tâches 61
II.4.3 Etapes 61
II.4.4 Activités 62

II.5 Discussion - Conclusion 64

CHAPITRE III. RECONNAISSANCE AUTOMATIQUE DE TACHES CHIRURGICALES 66

III.1 Indexation et Recherche de vidéos par le contenu (CBVR) 67
III.1.1 Caractérisation des vidéos 67
III.1.1.1 Analyse de la structure 68
III.1.1.1.1 Images 68
III.1.1.1.2 Sous-séquences 69
III.1.1.2 Extraction de caractéristiques visuelles 69
III.1.1.2.1 Caractéristiques statiques : couleurs, textures, formes 69
III.1.1.2.2 Caractéristiques dynamiques : Objets, Mouvement 70
III.1.1.3 Construction des signatures visuelles 71

III.1.2 Mesure de similitude 73
III.1.2.1 Alignement dynamique temporel (DTW) 73
III.1.2.2 Mesure de similitude de Piciarelli et al. 74
III.1.3 Catégorisation des vidéos 75
III.1.4 Synthèse 76

III.2 Caractérisation des vidéos : nos choix 77
III.2.1.1 Extraction de caractéristiques basées sur le mouvement 77
III.2.1.1.1 Histogrammes de mots visuels 77
III.2.1.1.2 Histogrammes de Mouvement 78
III.2.1.2 Normalisation des vidéos 79

III.3 Catégorisation des vidéos : nos choix 82
III.3.1 Mesure de similitude de Piciarelli et al. 82
III.3.2 Recherche des plus proches voisins 83

III.4 Evaluation 85
III.4.1 La base de sous-séquences vidéos 85
III.4.2 Mesure de la performance : l'aire sous la courbe ROC 85
III.4.3 Résultats 87
III.4.3.1 Mesure de similitude de Piciarelli et al. 87
III.4.3.2 Influence du choix de la caractérisation des vidéos 89
III.4.3.3 Temps de calcul 93

III.5 Discussion – Conclusion 94

CHAPITRE IV. SEQUENÇAGE MULTI-ECHELLES D'UNE VIDEO DE CHIRURGIE 96

IV.1 Modélisation du processus chirurgical 98
IV.1.1 Modèles Graphiques 99
IV.1.1.1 Les graphes quelconques 99
IV.1.1.2 Les arbres 100
IV.1.2 Information contextuelle en analyse de vidéos médicales 101
IV.1.2.1 Construction d'une chirurgie moyenne 101
IV.1.2.2 Modèles Statistiques utilisés en analyse de vidéos médicales 102
IV.1.2.2.1 Modèles Markoviens 102
IV.1.2.2.2 Les Champs Markoviens conditionnels 104
IV.1.2.2.3 Les Systèmes Linéaires Dynamiques 105
IV.1.2.3 Modélisation des relations entre les niveaux 105
IV.1.2.3.1 Arbres de décision 106
IV.1.2.3.2 Réseaux bayésiens 106
IV.1.3 Synthèse 107

IV.2 Construction d'arbres (Piciarelli et al.) 109
IV.2.1.1 Méthode de Piciarelli et al. 109
IV.2.2 Construction de l'arbre 111
IV.2.2.1 Méthode non supervisée 111
IV.2.2.2 Méthode supervisée 112
IV.2.3 Inférence de l'arbre 113

IV.2.4 Résultats 114
IV.2.4.1 Méthode non supervisée 114
IV.2.4.2 Méthode supervisée 115
IV.2.5 Synthèse 117

IV.3 Modélisation statistique multi-échelles 119
IV.3.1 Construction du modèle 119
IV.3.1.1 Réseau Bayésien 120
IV.3.1.2 HMM 123
IV.3.1.1 CRF 123
IV.3.1.2 Retour du HMM vers le réseau bayésien 125
IV.3.2 Caractérisation de la vidéo 126
IV.3.2.1 Structure de l'analyse 126
IV.3.2.2 Génération des observations 126
IV.3.2.2.1 Présence des instruments 127
IV.3.2.2.2 Analyse du mouvement 127
IV.3.3 Evaluation 128
IV.3.3.1 Optimisation des paramètres 129
IV.3.3.1.1 Utilisation des HMM 129
IV.3.3.1.2 Utilisation des CRF 131
IV.3.3.2 Inférence 132
IV.3.3.2.1 Réseau bayésien seul 132
IV.3.3.2.2 HMM 133
IV.3.3.2.3 CRF 133
IV.3.3.3 Résultats 134
IV.3.3.3.1 Réseau Bayésien 134
IV.3.3.3.2 Réseau Bayésien et HMM 135
IV.3.3.3.3 Réseau Bayésien et retour HMM « phases » 138
IV.3.3.3.4 Réseau Bayésien et CRF 139
IV.3.3.4 Conclusion 140

IV.4 Discussion – Conclusion 142

CHAPITRE V. DISCUSSION GENERALE 144

V.1 Description du processus chirurgical 144

V.2 Recherche des cas les plus proches 145

V.3 Modélisation statistique du processus chirurgical 146

CONCLUSION 148

BIBLIOGRAPHIE 152

ANNEXES 160

Table des figures

Figure 1. Principe de la fouille de données 24
Figure 2. Principe du raisonnement à base de cas 25
Figure 3. Principe de la CBIR 26
Figure 4. Principe de la CBVR 27
Figure 5. Les 3 axes du LaTIM 29
Figure 6. Principe des méthodes de reconnaissance automatique de tâches 38
Figure 7. Exemple de suivi de région entre deux images consécutives [48] 38
Figure 8. A gauche, extraction des vecteurs de caractéristiques des sous-séquences, permettant de découper automatiquement un geste chirurgical en mouvements élémentaires [49]; A droite, en bleu, le champ de mouvements approché par des polynômes spatiotemporels, en vert les champs de mouvements entre deux images consécutives mesurés par l'algorithme de Farnebäck [50] 39
Figure 9. A gauche, le principe de la méthode de segmentation et de catégorisation des sous-séquences proposée par G. Quellec et al. [51], à droite, la méthode de caractérisation associée40
Figure 10. Principe des méthodes de séquençage automatique des vidéos de chirurgies en tâches chirurgicales développées par l'équipe GD2MP 43
Figure 11. Scénario envisagé pour la reconnaissance du geste chirurgical 44
Figure 12. Différents niveaux de granularités que l'on peut trouver dans la littérature 47
Figure 13. Structure anatomique de l'œil humain 53
Figure 14. Les neufs tâches chirurgicales principales 55
Figure 15. Les trois nouveaux niveaux de granularités 57
Figure 16. Les phases de la chirurgie de la cataracte 58
Figure 17. Relations entre les différentes étapes et phases qu'il est possible de rencontrer dans une chirurgie de la cataracte selon la description de David Martiano et moi-même 58
Figure 18. Processus de construction de graphes pour la modélisation du processus chirurgical 60
Figure 19. Diagramme de transition des phases chirurgicales 61
Figure 20. Diagramme de transition des tâches chirurgicales 61
Figure 21. Zoom sur l'organisation des étapes appartenant à la tâche phacoémulsification du diagramme de transition des étapes chirurgicales 62
Figure 22. Zoom sur l'organisation des activités appartenant à la tâche phacoémulsification du diagramme de transition des activités chirurgicales 63
Figure 23. Caractérisation des vidéos par leur contenu visuel 68
Figure 24. Différentes façon d'utiliser la structure de la vidéo pour l'extraction des caractéristiques visuelles 68
Figure 25. Méthode de caractérisation d'une image par sacs de mots visuels ; en haut, la construction d'un dictionnaire de mots visuels à partir de la base d'apprentissage ; en bas, la caractérisation d'une nouvelle image à partir d'un sac de mots visuels 72

Figure 26. Principe de l'algorithme DTW ; à gauche, l'alignement de deux séquences réalisé avec l'algorithme DTW ; à droite, le chemin de moindre coût avec en vert un exemple d'enveloppe de restriction de la recherche 73

Figure 27. Partitionnement de trajectoires selon la méthode de Piciarelli et al. [44]..... 74

Figure 28. Principe de la mesure de similitude proposée par Piciarelli et al. [44] 75

Figure 29. Exemple de points d'intérêts spatiotemporels STIP extraits dans deux images de vidéos de chirurgies de la cataracte..... 78

Figure 30. Différentes zones de l'œil visualisées dans le champ de vue de la caméra ; en rose, la zone d'action principale des outils..... 80

Figure 31. Normalisation des vidéos de chirurgie de la cataracte à partir de la détection du centre (de la pupille et de l'iris) et du facteur d'échelle 81

Figure 32. Principe de la méthode de reconnaissance automatique des tâches chirurgicales 84

Figure 33. Présentation de la courbe ROC..... 86

Figure 34. Courbes ROC obtenues sans la normalisation des vidéos pour chaque tâche chirurgicale avec les signatures en histogrammes de mouvement (HM)..... 90

Figure 35. Courbes ROC obtenues avec la normalisation (Recalage + Sélection d'une ROI + Mise à l'échelle) des vidéos pour chaque tâche chirurgicale avec les signatures en histogrammes de mouvement (HM)..... 90

Figure 36. Courbes ROC obtenues sans normalisation des vidéos pour chaque tâche chirurgicale avec les signatures en histogrammes de mots visuels (BoW) construit à partir des descripteurs STIP..... 92

Figure 37. Courbes ROC obtenues avec la normalisation des vidéos (sélection d'une ROI) pour chaque tâche chirurgicale avec les signatures en histogrammes de mots visuels (BoW) construit à partir des descripteurs STIP 92

Figure 38. Principe général de notre méthode de séquençage automatique d'une chirurgie requête..... 98

Figure 39. A gauche un exemple de graphe non orienté acyclique; A droite, un exemple de cycle..... 99

Figure 40. A gauche un exemple de graphe non orienté; A droite, un exemple de graphe orienté..... 100

Figure 41. Exemple d'arbre 100

Figure 42. Exemple de segmentation d'une chirurgie requête en la recalant sur une chirurgie moyenne via l'algorithme DTW..... 101

Figure 43. Exemple de diagramme de changement d'état à 4 états 102

Figure 44. Exemple de modèle de Markov à 4 états cachés ; les flèches en pointillés indiquent les sorties probables à chaque passage dans un état..... 103

Figure 45. A gauche, exemple de structure d'une chaîne de Markov cachée simple ; à droite, exemple de structure en chaîne d'un CRF 104

Figure 46. Exemple d'un DAG d'un réseau bayésien modélisant 3 niveaux de granularité ; les nœuds jaunes représentent les activités, les nœuds bleus les étapes et les nœuds verts les phases 107

Figure 47. Mise à jour de l'arbre pour l'acquisition d'une nouvelle trajectoire [44]..... 110

Figure 48. Processus de construction de l'arbre 110
Figure 49. Méthode de construction supervisée d'arbres 112
Figure 50. Parcours de l'arbre pour la reconnaissance automatique de tâches chirurgicales 113
Figure 51. Exemples de comparaisons de vidéos 2 à 2 ; en haut, l'évolution de la mesure de distance entre les deux vidéos ; en bas, la comparaison de l'enchaînement des phases chirurgicales entre la vidéo requête (en haut) et la vidéo de référence (en bas) 115
Figure 52. Exemple d'arbre obtenu avec 11 vidéos 117
Figure 53. Principe général de la modélisation multi-échelles de la chirurgie, combinant un réseau Bayésien (à gauche) et 3 modèles de Markov (à droite) 119
Figure 54. Méthodologie du système d'annotation automatique 120
Figure 55. Méthodologie du système d'annotation automatique pour lequel les HMM ont été remplacés par des CRF pour la modélisation du déroulement temporel de la chirurgie 124
Figure 56. Exemple de réseau bayésien à deux niveaux de description, en violet les nœuds d'observation ; Les nœuds O7 à O11 permettent d'apporter les informations obtenus suite à l'inférence du HMM 125
Figure 57. Méthodologie du système d'annotation automatique, en retournant les valeurs issues du HMM « phases » 126
Figure 58. Découpage de la vidéo requête en en sous-séquences de taille fixe 126
Figure 59. Exemple de réseau bayésien à deux niveaux de description, en violet les nœuds d'observation 127
Figure 60. Utilisation d'une validation croisée pour l'évaluation du modèle 129

Table des figures

Table des tableaux

Tableau 1. Tableau récapitulatif des différentes méthodes proposées en analyse automatique de vidéos dans le domaine médical 35

Tableau 2. Bases de données utilisées pour les différents travaux en analyse automatiques de vidéos médicales 50

Tableau 3. Statistiques de la base de séquences vidéo utilisées pour évaluer la méthode de reconnaissance automatique de tâches chirurgicales (base de test) 85

Tableau 4. Possibilités de résultats d'un classifieur binaire pathologique/non pathologique 86

Tableau 5. Présentation des résultats obtenus avec la distance de Piciarelli et al. pour les quatre types de distances locales ; comparaison avec l'algorithme DTW associé avec la distance Bhattacharyya 87

Tableau 6. résultats obtenues avec la distance de Piciarelli et al., avec la distance locale de Bhattacharyya, pour différents choix de facteur de grandissement de la fenêtre glissante 88

Tableau 7. Résultats en termes d'aire $\boldsymbol{Az}$ sous la courbe ROC obtenus avec les signatures visuelles en histogrammes de mouvement (HM) construites à partir du flux optique 89

Tableau 8. Résultats en termes d'aire $\boldsymbol{Az}$ sous la courbe ROC obtenus avec les signatures en histogrammes de mots visuels (BoW) construit à partir des descripteurs STIP 91

Tableau 9. Temps de calcul en secondes pour les différents éléments de la caractérisation des vidéos 93

Tableau 10. Résultats obtenus pour deux vidéos de test après inférence de l'arbre 116

Tableau 11. Résultats obtenus pour deux vidéos de test après inférence de l'arbre (signatures globales) 116

Tableau 12. Paramètres optimaux obtenus pour l'utilisation de la présence des instruments comme observations 129

Tableau 13. Paramètres optimaux obtenus pour l'utilisation de l'analyse du mouvement dans la vidéo comme source d'observations 130

Tableau 14. Paramètres optimaux obtenus pour le retour des résultats de l'inférence du HMM « phases » dans le réseau bayésien 130

Tableau 15. Paramètres optimaux obtenus pour l'utilisation de l'analyse du mouvement comme source d'observations avec des histogrammes de mouvement comme signatures visuelles 131

Tableau 16. Paramètres optimaux obtenus dans le cadre de l'utilisation de l'information de présence des instruments comme source d'observations 131

Tableau 17. Comparaison des méthodes d'élagage du réseau bayésien 135

Tableau 18. Evaluation du modèle combinant un réseau bayésien et des HMM, avec comme source d'observations la présence des instruments dans le champ de vue de la caméra 135

Tableau 19. Evaluation du modèle combinant un réseau bayésien et des HMM, avec comme source d'observations les résultats de la recherche des cas les plus proches (comparaison d'histogrammes de mouvement) 136

Tableau 20. Evaluation du modèle combinant un réseau bayésien et des HMM, avec comme source d'observations les résultats de la recherche des cas les plus proches (comparaison d'histogrammes de mouvement) ; Comparaison avec et sans normalisation des vidéo (recalage spatial, sélection d'une ROI et mise à l'échelle)..... 137

Tableau 21. Evaluation du modèle combinant un réseau bayésien et des HMM, avec comme source d'observations les résultats de la recherche des cas les plus proches (comparaison d'histogrammes de mots visuels) 137

Tableau 22. Comparaison du modèle combinant un réseau bayésien et des HMM, avec retour du résultat de l'inférence du HMM pour le niveau « phases » vers le réseau bayésien ; utilisation de la présence des instruments comme source d'observations..... 138

Tableau 23. Comparaison du modèle combinant un réseau bayésien et des HMM, avec retour du résultat de l'inférence du HMM pour le niveau de description en phases vers le réseau bayésien ; utilisation des résultats de la recherche des cas les plus proches comme source d'observations (comparaison d'histogrammes de mouvement)..... 139

Tableau 24. Comparaison du modèle combinant un réseau bayésien et des HMM avec le modèle combinant un réseau bayésien et des CRF, avec comme source d'observations la présence des instruments, puis les résultats de la recherche des cas les plus proches (comparaison d'histogrammes de mouvement)..... 139

Glossaire

ACC : Analyse Canonique des Corrélations (ou CCA « canonical-correlation analysis » en anglais), permet de comparer ces deux groupes de variables pour savoir si ils décrivent un même phénomène

ACP : Analyse en Composantes Principales (ou PCA en anglais pour « Principal Component Analysis ») , permet de réaliser un changement de base de la base des variables initiale vers une base de variables non corrélées (composantes principales)

ANN : Approximate Nearest Neighbor Searching, méthode approximative de recherche de plus proches voisins

BEMD : Bidimensional Empirical Mode Decomposition (Décomposition en modes empiriques bidimensionnelle)

BoW : Bag of visual Word (sacs de mots visuels)

CHRU : Centre Hospitalier et Régional Universitaire

CBIR : Content Based Image Retrieval (recherche d'images par le contenu)

CBR/RàPC/RBC : Case-Based Reasoning/Raisonnement à Partir de Cas/ Raisonnement à Base de Cas

CBVR : Content Based Video Retrieval (recherche de vidéos par le contenu)

CRF : Conditional Random Fieds (Champs Markoviens conditionnels)

DAG : Directed Acyclic Graph (graphes acycliques orientés)

DMP : Dossier Médical Personnel

DSmT : Théorie de Dezert-Smarandache, combinaison formelle de n'importe quel type (certain, incertain, paradoxal) d'information

DTW : Dynamique Time Warping (alignement dynamique temporel), mesure de similitude entre deux séquences temporelles dont la durée ou la vitesse peuvent varier

EFDTW : Extented Fast Dynamic Time Warping, combinaison entre l'algorithme avec la mesure de distance EMD

EMD : Earth Mover's Distance (distance du cantonnier), mesure de distance entre deux vecteurs, évalue la quantité de travail nécessaire pour modifier un vecteur en un autre

FDTW : Fast Dynamique Time Warping, extension plus rapide de l'algorithme DTW

FN : Faux Négatifs, représente le nombre d'éléments positifs détectés négatif

FP : Faux Positifs, représente le nombre d'éléments négatif détectés positifs

GD2MP : Equipe « Gestion des données médicales multimodales partagées pour l'aide à la décision » au sein du LaTIM

Granularité : niveau de description (d'une description fine à une description grossière)

HMM : Hidden Markov Model (modèle de Markov Cachée)

HOF : Histogram of Optical Flow, histogramme construit à partir des informations issues du flux optique

HOG : Histogram of Oriented Gradients, histogrammes de répartition des intensités des gradients

HSV : Hue Saturation Value (Teinte, Saturation, Valeur), autre type de codage des couleurs

INSERM : Institut National de la Santé Et de la Recherche Médicale

IRM : Imagerie par Résonance Magnétique

LaTIM : Laboratoire de Traitement de l'Information Médicale (INSERM – UMR 1101)

LBP : Local Binary Pattern (motifs binaires locaux), caractéristiques utilisées pour reconnaître les textures dans une image

LDS : Linear Dynamical System (systèmes linéaires dynamiques)

MCTM : Markov Clustering Topic Model

MMC : modèle de Markov Cachée ou HMM pour « Hidden Markov Model » en anglais

MPEG : Moving Picture Experts Group en anglais, ici : format de codage de vidéos

RFID : radio frequency identification (radio-étiquettes)

RGB : Red, Green, Blue (Rouge, Vert, Bleu), il s'agit du système le plus simple et le plus courant de représentation des couleurs en informatique

ROC : Receiver Operating Characteristic, mesure la performance d'un classifieur binaire

ROI : Region Of Interest (région d'intérêt)

SIFT : Scale-Invariant Feature Transform (transformation de caractéristiques visuelles invariante à l'échelle), descripteur local qui permet de représenter le contenu visuel d'une image

SVM : Support Vector Machines (machine à vecteurs de support)

UMR : Unité Médicale de Recherche

VN : vrais négatifs, représente le nombre d'éléments négatif détectés négatif

VP : Vrais Positifs, représente le nombre d'éléments positifs détectés positifs

YUV : type de codage de couleurs ; Y représente la luminance et U et V la chrominance

Introduction

Le domaine médical génère des données en grande quantité dont l'exploitation peut être un véritable atout pour l'aide à la pratique médicale. La formation médicale se faisant majoritairement par l'apprentissage et l'expérience, ces données ont un réel intérêt de par la quantité d'information qu'elles contiennent : des exemples, des diagnostics posés par des médecins plus expérimentés, des cas similaires, etc… Mais elles ne sont pas toujours facilement consultables par les médecins bien que les ressources informatiques actuelles offrent aujourd'hui les capacités de stockage et de traitement suffisantes pour automatiser l'exploration et le traitement de ces données. Cela ouvre la voie vers une aide à la prise de décision, non seulement en facilitant la consultation des cas similaires, mais également en émettant automatiquement des hypothèses de diagnostics et des recommandations.

Dans ce domaine, le LaTIM a développé des approches originales d'aide à la prise de décision dans le domaine de l'ophtalmologie, en collaboration avec le CHRU de Brest. Les méthodes développées, notamment pour l'aide au dépistage de la rétinopathie diabétique, ont l'originalité de ne pas s'appuyer sur des méthodes de segmentation classiques, pour détecter les lésions typiques de cette pathologie. L'approche adoptée consiste à s'appuyer sur des méthodes de recherche de cas similaires par le contenu, en utilisant à la fois les images et les informations sémantiques contextuelles telles que l'âge, le sexe ou les antécédents du patient. L'aide à la prise de décision ne se limite pas à l'analyse automatique de clichés médicaux (images de fond de l'œil, images IRM, TEP, scanner, etc…). Les chirurgies sous contrôle vidéo, tels que la chirurgie de la cataracte, fournissent également une quantité importante de données médicales encore très peu exploitées. Ces données peuvent être utilisées, tout comme pour l'aide au diagnostic, pour l'aide à la pratique chirurgicale. Les informations contenues dans les cas archivés peuvent s'avérer être une aide précieuse, pour les jeunes chirurgiens en apprentissage ou dans le cadre de la chirurgie robotisée, en permettant le contrôle du déroulement des actes chirurgicaux. En effet, nous pouvons imaginer utiliser les méthodes de recherche de cas similaires par le contenu pour reconnaître automatiquement le geste chirurgical, présenter des exemples ciblés de pratiques de chirurgiens plus expérimentés, détecter des situations anormales ou à risque et générer des alertes et des recommandations adaptées.

L'objectif de cette thèse est, dans la lignée des méthodes déjà implémentées par le LaTIM, d'exploiter les données vidéo enregistrées lors de chirurgies, pour apporter une aide en temps réel aux chirurgiens. De par l'inexistence d'une base de données de vidéos chirurgicales commune, les méthodes développées dans cette thèse s'appliquent sur une base de cas de chirurgie de la cataracte, collectée par le LaTIM. Le principal défi de ce travail est d'être capable d'analyser en temps réel le processus chirurgical et de reconnaître à tous moment le geste chirurgical effectué. Pour cela, différents défis se présentent à nous : tels que la conceptualisation du processus chirurgical. Nous cherchons à considérer la chirurgie comme une succession de tâches, de phases ou de gestes chirurgicaux. Il est alors nécessaire de trouver une description qui permet d'analyser le processus chirurgical avec suffisamment de précision et une reconnaissance aisée. Un autre défi de ce travail de thèse est que les méthodes d'analyse proposées doivent gérer des vidéos de chirurgies effectuées par différents chirurgiens, plus ou moins expérimentés et enregistrées par des

systèmes d'acquisition variables. Il existe dans la littérature, très peu de méthodes permettant d'analyser des vidéos de chirurgies en temps réel à un niveau de description suffisamment fin.

Nous proposons dans cette thèse un certain nombre de réponses à ces différentes problématiques. Différentes réflexions ont été menées autour de la caractérisation des vidéos par leur contenu visuel, la recherche des cas les plus proches dans la base de données, mais également la modélisation statistique du processus chirurgical. Nous avons ainsi tout d'abord mené un travail de réflexion avec le service d'ophtalmologie du CHRU de Brest, pour définir une nouvelle description multi-échelles de la chirurgie de la cataracte, afin d'analyser la chirurgie avec une plus grande précision. Les vidéos de chirurgie de la base de données ont été annotées manuellement par les chirurgiens qui ont déterminé les séquençages temporels des vidéos pour chacune des descriptions afin de créer une base de référence. Nous avons également travaillé à la reconnaissance des tâches chirurgicales en ramenant tout d'abord le problème à une catégorisation de séquences vidéo, où chaque séquence représente une des tâches chirurgicales, exécutées au cours d'une intervention. En nous inspirant des méthodes déjà développées par le LaTIM, notamment en recherche d'images par le contenu, nous nous sommes tout d'abord orientés vers une méthode de reconnaissance automatique du geste chirurgical via une recherche des cas les plus proches dans la base de données. Les méthodes de recherche de vidéos par le contenu, permettent de reconnaître le geste le plus probable, en fonction des cas les plus proches retrouvés. Nous nous sommes ensuite attelés au séquençage automatique d'une vidéo de chirurgie complète en gestes chirurgicaux. Nous cherchons alors à assigner un label (le geste chirurgical le plus probable) à chaque image de la vidéo. L'utilisation de notre connaissance du déroulement de la chirurgie, en plus du contenu visuel de la vidéo, nous a semblée une approche intéressante pour apporter une information contextuelle lors du choix du label. Une réflexion a été menée sur la construction d'un modèle statistique du processus chirurgical, appris à partir de la connaissance des cas précédemment archivés. Cela permet d'apporter une information contextuelle lors de la reconnaissance automatique du geste chirurgical.

Le manuscrit est organisé en quatre chapitres. Le premier chapitre présente le contexte du travail de thèse autour de la réutilisation des archives médicales numériques et de l'analyse automatique de vidéos médicales. Le second chapitre présente le travail réalisé en collaboration avec le service d'ophtalmologie du CHRU de Brest, pour construire la nouvelle description multi-échelles de la chirurgie de la cataracte. Les diagrammes de transition obtenus pour chacun des niveaux sont également présentés. Ces diagrammes de transition serviront de base à la construction des modèles statistiques du processus chirurgical présentés dans le Chapitre IV. Dans le chapitre III, nous présentons l'état de l'art sur la recherche de vidéos par le contenu, et les choix que nous avons faits en termes de caractérisation et de catégorisation des vidéos. Nous explicitons la méthode que nous avons choisie pour la reconnaissance automatique des tâches chirurgicales basée sur la recherche de vidéos similaires par le contenu. Les performances de classification de la méthode sont ensuite évaluées sur une base de séquences vidéo, où une séquence représente une tâche chirurgicale. Dans le chapitre IV, nous introduisons la modélisation du processus chirurgical pour le séquençage multi-échelle des vidéos. Nous présentons des méthodes de modélisation statistique de l'état de l'art. Les deux méthodes choisies et implémentées sont détaillées puis évaluées. Dans le Chapitre V, les principales réponses apportées dans cette thèse pour l'analyse automatique de vidéos chirurgicales seront discutées.

Chapitre I. CONTEXTE

CHAPITRE I. CONTEXTE 22

I.1 Réutilisation des archives médicales numériques 23
I.1.1 Les archives médicales numériques 23
I.1.2 La fouille de données 24
I.1.2.1 Le raisonnement à partir de cas 25
I.1.2.2 La Recherche d'images par le contenu (CBIR) 26
I.1.2.3 La recherche de vidéos par le contenu (CBVR) 27
I.1.3 Positionnement du travail de thèse dans les recherches du LaTIM 28
I.1.3.1 La Recherche d'images par le contenu (CBIR) 29
I.1.3.2 L'analyse automatique de vidéos chirurgicales 30
I.1.4 Conclusion 30

I.2 Aide à la chirurgie en temps réel 31
I.2.1 Travaux principaux en analyse automatique de vidéos 31
I.2.1.1 Dans le domaine médical 31
I.2.1.1.1 Résumé automatique d'examens 31
I.2.1.1.2 Annotation automatique de vidéos 32
I.2.1.1.3 Recherche de cas similaires 33
I.2.1.1.4 Evaluation des compétences des chirurgiens 33
I.2.1.1.5 L'analyse de scènes opératoires 34
I.2.1.1.6 Conclusion 34
I.2.1.2 Analyse de séquences vidéos dans d'autres domaines 36
I.2.1.3 Conclusion 37
I.2.2 Les travaux du LaTIM dans l'analyse automatique de vidéos chirurgicales 37
I.2.2.1 Reconnaissance automatique de tâches chirurgicales 37
I.2.2.2 Séquençage automatique de vidéos de chirurgies 40
I.2.3 L'analyse automatique de vidéos en temps réel 41
I.2.3.1 Algorithmes rapides 41
I.2.3.2 Algorithmes d'analyse « en direct » 41
I.2.3.3 Conclusion 42
I.2.4 Scénario envisagé 42
I.2.4.1 Reconnaissance du geste chirurgical 42
I.2.4.2 Détection d'événements anormaux et génération d'alertes 44

I.3 Discussion - Conclusion 45

I.1 Réutilisation des archives médicales numériques

Des premiers ordinateurs qui ne disposaient que de quelques dizaines de kilo-octets de mémoire de masse, aux Data Centers d'aujourd'hui où sont stockés des peta-octets de données numériques[1], l'humanité a franchi un cap historique. A l'usage de l'informatique pour le calcul scientifique, la gestion informatique de processus, les systèmes d'information, s'est rajoutée l'exploitation des données massives ("Big Data") pour extraire des connaissances, construire de l'expertise, directement à partir des données numériques archivées dans tous les domaines : réseaux sociaux, économie, finance, écologie, cartographie, multimédia, …. La santé est sûrement un des domaines qui va bénéficier le plus de l'exploitation de la quantité de plus en plus grande de données qui sont enregistrées chaque jour dans les services hospitaliers, les cabinets médicaux, chez les professionnels de santé. L'exploitation informatique de données épidémiologiques a déjà permis de nombreuses avancées en termes d'amélioration de la santé des populations. L'exploitation de données personnelles, mais obligatoirement anonymisées, doit permettre d'aller encore plus loin au bénéfice des patients. La gestion et l'utilisation des données massives est donc un véritable enjeu et le domaine médical n'échappe pas à la règle. Outre le problème de gestion et de sécurisation de ces données, elles présentent un réel potentiel pour faciliter le travail des médecins, notamment par la mise en place d'outils d'aide à la décision.

I.1.1 Les archives médicales numériques

Comme dans de nombreux domaines, une grande quantité de données médicales est maintenant sauvegardée numériquement. En France, par exemple, le Dossier Médical Personnel numérique (DMP) permet à chaque français qui le souhaite d'avoir un dossier médical numérique. Il contient des informations médicales telles que les antécédents médicaux, des images, des résultats d'analyses en provenance de différents professionnels de santé. Au 1er aout 2015, plus de 549 940 dossiers avaient déjà été créés[2]. Le volume de données archivées au sein des différents services des centres hospitaliers ou des centres de dépistage ne cesse de croitre. Il s'agit de données textuelles, tel que l'âge, le sexe, les antécédents médicaux des patients, des mesures biologiques, des signaux mono-multi-dimensionnels (ECG, EEG, électromyogrammes, …), des images (radiographies, scanners, IRM, fond de l'œil…), ou des données vidéo dans le cas des examens endoscopiques, des échographies ou des chirurgies sous contrôle vidéo par exemple. Ces données sont archivées et peuvent être consultées par des médecins lorsque cela est nécessaire, de la même manière qu'ils consulteraient un ouvrage médical, ou des articles de référence, en cas de difficulté pour poser un diagnostic, ou simplement pour le conforter. Les archives numériques vont pouvoir regrouper un très grand nombre de cas cliniques et être des sources d'information très riches, mais cependant difficilement exploitables directement par les médecins. Il peut être effectivement très long et complexe de parcourir l'ensemble des cas archivés. Or les technologies numériques offrent de nombreux outils pour extraire de l'information des bases de données, en

1 http://www-935.ibm.com/services/ca/en/it-services/datacenter-cldc.html

2 www.dmp.gouv.fr

cherchant par exemple des cas similaires. C'est l'approche que nous développons dans notre travail de thèse, appliquée au suivi de procédures chirurgicales.

I.1.2 La fouille de données

La fouille de données ou exploration de données (« Data mining » en anglais) permet d'extraire des connaissances à partir de données. Comme cela est présenté dans la Figure 1, ce domaine est à l'intersection des statistiques, de l'apprentissage automatique et des bases de données. Il regroupe l'ensemble des méthodes permettant le passage des données brutes à la connaissance d'un domaine. Différentes approches sont possibles, la première consiste à *décrire* automatiquement les données brutes. C'est le cas par exemple des méthodes de regroupement (« clustering » en anglais) qui cherchent à rassembler automatiquement les données dans des groupes distincts. Un deuxième type d'approches consiste à utiliser les données et les connaissances qui leurs sont associées pour *prédire* ou expliquer un ou plusieurs phénomènes observables et effectivement mesurés (ensemble de tests). Ces méthodes s'appuient le plus souvent sur l'utilisation de méthodes statistiques telles que des arbres de décision, des réseaux bayésiens ou le raisonnement à base de cas.

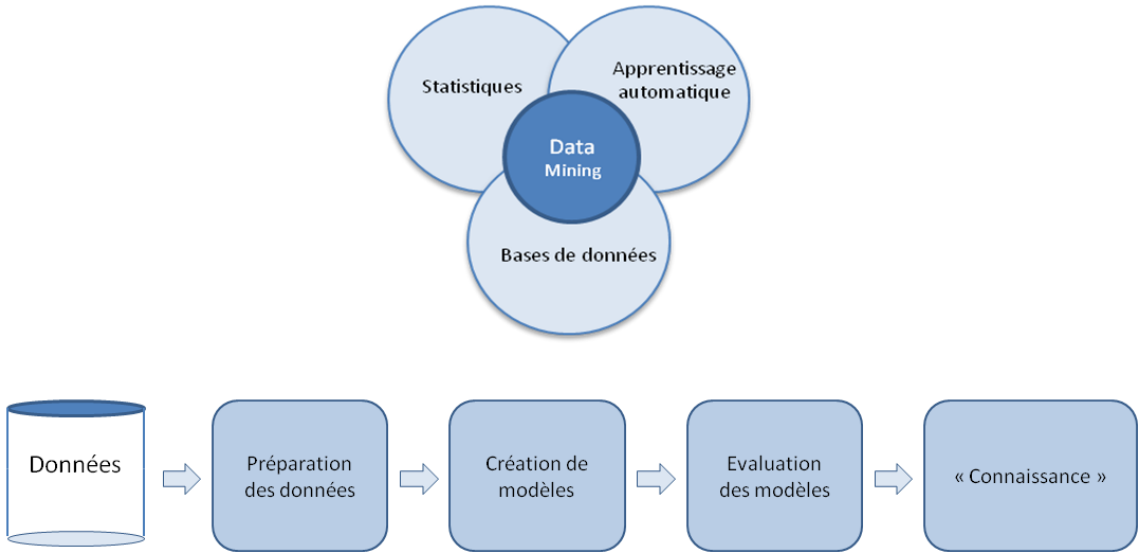


Figure 1. Principe de la fouille de données

Les applications de la fouille de données sont très vastes [1]. Il s'agit de méthodes très utilisées, indispensables aujourd'hui du fait de la grande quantité de données stockées et des ressources disponibles pour les analyser et les exploiter automatiquement. Ces méthodes sont par exemple de plus en plus utilisées dans le domaine de l'industrie pour la maintenance préventive et l'optimisation, voire la détection de fraudes pour protéger la propriété intellectuelle et le secret industriel. Dans le domaine médical, la fouille de données est également très largement utilisée, essentiellement dans le cadre de l'épidémiologie, c'est-à-dire l'étude des facteurs influant sur la santé et les maladies de groupes de population. Les méthodes prédictives de fouille de données trouvent également tout leur intérêt dans le domaine médical, dans le cadre de l'aide à la décision.

C'est le cas des méthodes de raisonnement à base de cas, particulièrement adaptées à l'utilisation des archives médicales numériques.

I.1.2.1 Le raisonnement à partir de cas

Le raisonnement à partir de cas (RàPC) appartient à la classe des méthodes prédictives de fouille de données et apportent une première réponse pour l'extraction automatique de connaissances à partir d'une base de cas. Il s'agit d'une approche intuitive de recherche des cas les plus proches, calquée sur le comportement humain. Aamodt et al. présentent les différents aspects et approches du raisonnement à partir de cas [2]. Le principe est présenté dans la Figure 2. Lorsqu'un nouveau cas est présenté en requête, il est comparé aux cas archivés dans la base de données (*Recherche*). Pour cela, il est nécessaire de représenter les différents cas par des descripteurs facilement utilisables par un ordinateur. Généralement un vecteur de caractéristiques est utilisé (*Représentation*). Chaque cas de la base a été préalablement annoté par un professionnel qui à chaque cas associe sa solution. La connaissance de l'annotation associée aux cas les plus proches sélectionnés automatiquement au sein de la base va permettre de classifier (diagnostiquer) le cas placé en requête (*Réutilisation*). La solution choisie est alors testée : si le diagnostic n'est pas correct, il est corrigé (*Révision*). Lorsque la solution est validée, elle est conservée et ajoutée à la base de données (*Conservation*).

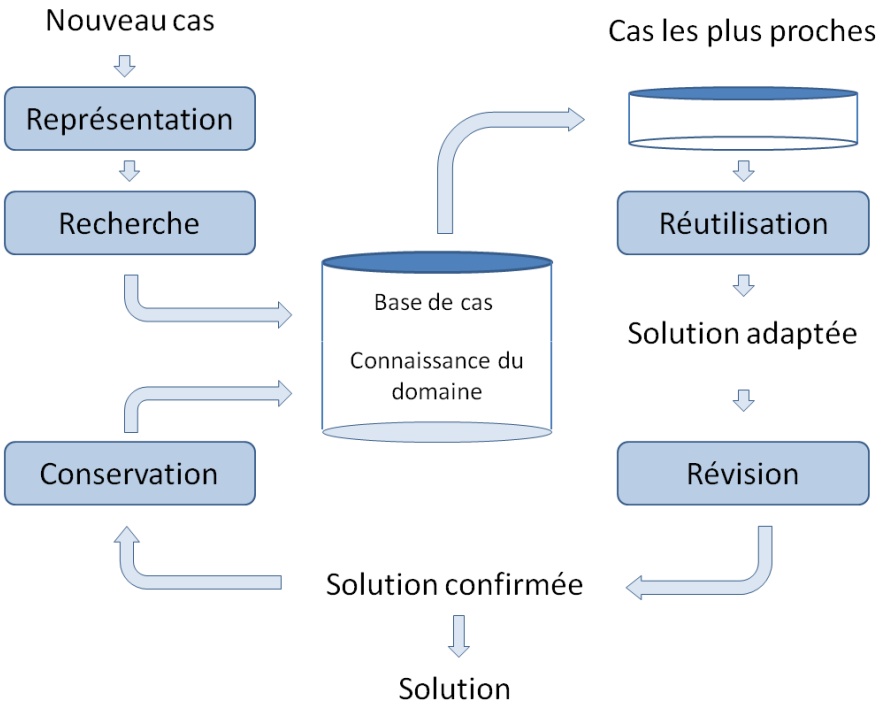


Figure 2. Principe du raisonnement à base de cas

Plusieurs points clés apparaissent dans ce système pour que ce dernier soit efficace et rapide. Une recherche de plus proches voisins pertinente est fortement liée au choix du vecteur de caractéristiques (descripteur) et de la mesure de similitude choisie pour comparer deux cas. De plus, la

construction d'une base de cas initiale et le choix des annotations (classes) associées aux données sont également à ne pas négliger. La base de cas initiale doit couvrir tout ou une grande partie de l'ensemble d'application afin d'être représentative des différents cas qui pourront être rencontrés.

Le raisonnement à partir de cas est général et peut être appliqué à une grande variété de domaines et de données. Il existe des méthodes plus spécifiques qui se focalisent sur un type de données, telles que les méthodes de recherche d'images ou de vidéos par le contenu.

I.1.2.2 La Recherche d'images par le contenu (CBIR)

Les méthodes de recherche d'images par le contenu (CBIR pour « Content Based Image Retrieval » en anglais) se basent sur des principes similaires (Figure 3) à ceux des méthodes RàPC, et peuvent être un des éléments des systèmes RàPC. De même que pour les méthodes de raisonnement à partir de cas, à chaque image de la base est associé un descripteur (que l'on appellera signature visuelle) et éventuellement une classe [3]. L'image requête est comparée via sa signature au cas de la base de données afin d'en trouver les cas les plus proches. La spécificité de ces méthodes est qu'elles se basent uniquement sur la comparaison du contenu visuel des images, sans nécessairement intégrer des aspects sémantiques. Les signatures peuvent donc ne contenir que des caractéristiques de bas niveaux (proches du signal) telles que des informations sur la couleur, les formes ou la texture par exemple (*Caractérisation*).

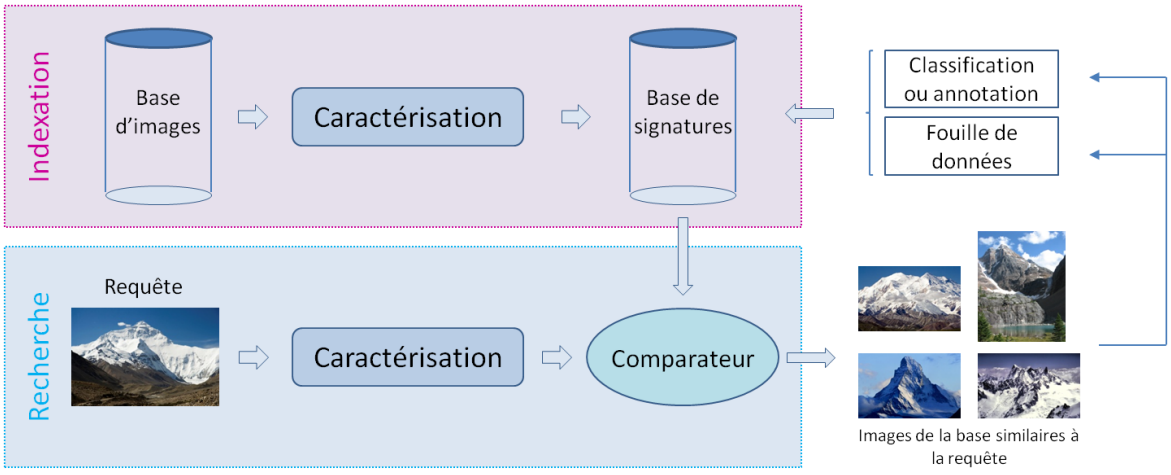


Figure 3. Principe de la CBIR

Ces méthodes trouvent leurs applications dans de nombreux domaines tels que la reconnaissance faciale, le filtrage des images pornographiques ou pédopornographiques, la lutte contre la contrefaçon, la cartographie ou l'imagerie médicale. Dans le domaine de l'imagerie médicale, ces méthodes peuvent par exemple fournir rapidement des exemples de cas similaires aux médecins, et donc apporter une aide au diagnostic. En s'appuyant sur les cas les plus proches sélectionnés, des algorithmes d'aide à la décision peuvent aider à déterminer s'il s'agit d'un cas pathologique ou sain. Cela permet des applications dans des systèmes de dépistage pour limiter la charge des praticiens. C'est le cas du prototype teleOphta développé au sein de l'équipe GD2MP du LaTIM

(paragraphe I.1.3), qui permet le tri automatique des cas sains dans le cadre du dépistage de la rétinopathie diabétique [4]. Avec plus de précision encore, les méthodes de classification basées sur la recherche par le contenu de cas similaires peuvent aider au comptage et à la localisation de lésions ou de tumeurs, proposant ainsi une alternative aux méthodes classiques de segmentation. Une revue de littérature a été proposée par Müller et al. autour des applications de la CBIR dans le domaine médical [5].

I.1.2.3 La recherche de vidéos par le contenu (CBVR)

La problématique de notre travail de thèse est l'aide per-opératoire en chirurgie mini-invasive, sous microscope, avec enregistrement vidéo des interventions. Nous avons donc besoin d'analyser les vidéos, et notre approche se base sur la recherche de vidéos par le contenu dans des archives vidéo d'interventions. C'est un domaine de recherche récent en vision par ordinateur. Les chercheurs dans le domaine se sont d'abord penchés sur l'adaptation des méthodes de recherche d'images par le contenu à la recherche de vidéos. La recherche de vidéos par le contenu (CBVR pour « Content based Video Retrieval » en anglais) permet d'extraire des informations pertinentes sur des vidéos présentées en requête par un utilisateur en les comparant aux vidéos archivées dans une base de données.

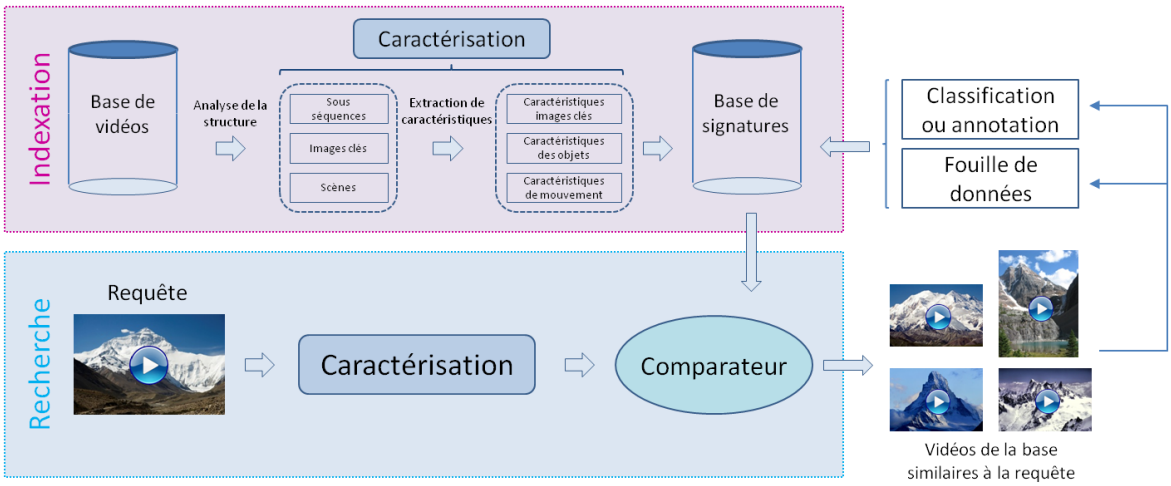


Figure 4. Principe de la CBVR

Weiming et al. proposent une revue de littérature des méthodes et applications de la recherche de vidéo par le contenu [6]. Une première idée intuitive consiste à utiliser telles quelles les méthodes de recherche d'images par le contenu, en considérant les vidéos comme une succession d'images. Cependant, on néglige alors l'aspect dynamique de la vidéo, pourtant essentiel. Ainsi, comme présenté dans la Figure 4, à la différence de la caractérisation d'images, la *caractérisation* des vidéos se fait en deux étapes clés. La première étape consiste à *analyser la structure* de la vidéo. On peut par exemple considérer la vidéo comme une succession d'images clés, de sous-séquences ou de scènes que l'on va ensuite caractériser par leur contenu visuel. *L'extraction de ces caractéristiques* peut être calquée sur les méthodes de CBIR, en utilisant les informations

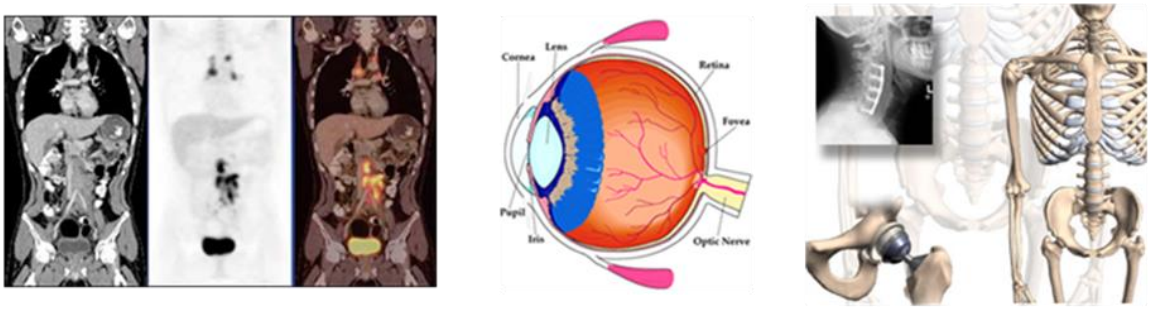
de couleurs, de textures ou de formes dans chacune des images. Cependant, d'autres opportunités s'offrent à nous avec la vidéo : l'extraction du mouvement entre plusieurs images consécutives peut en particulier apporter une information très pertinente. De même, plutôt que de se limiter aux informations de forme des objets contenus dans la vidéo, on peut également les suivre au cours du temps et utiliser, entre autre, les informations fournies par leurs trajectoires. La *comparaison* de séquences vidéo introduit également un autre aspect par rapport à la comparaison d'images. En plus des problèmes de mise à l'échelle spatiale, se pose des problèmes de mise à l'échelle temporelle. Deux séquences vidéo sémantiquement similaires peuvent en effet avoir des durées différentes ou représenter des actions dont les vitesses d'exécution sont différentes, du fait de l'expertise du chirurgien ou des aléas d'intervention. La mesure de similitude utilisée devra alors être à même de gérer ces distorsions temporelles.

Les méthodes de CBVR permettent de rechercher des vidéos similaires, de les classifier, mais aussi de les résumer ou de les segmenter automatiquement en scènes ou actions d'intérêt. Ces techniques permettent ainsi de développer des méthodes d'analyse automatique de vidéos variées et trouvent particulièrement leur place dans l'analyse automatique de vidéos issues de la vidéosurveillance. Ce domaine fournit une grande quantité de données qu'il est difficile d'exploiter manuellement en temps réel. Mais ces méthodes de recherche de vidéos trouvent également leur utilité dans l'analyse automatique et rapide de vidéos sportives et, plus récemment, dans le domaine médical. En effet, ces méthodes sont particulièrement adaptées dans le cadre des examens endoscopiques (coloscopie, bronchoscopie…), des chirurgies sous contrôle vidéo telles que les chirurgies endoscopiques (la chirurgie laparoscopique par exemple) et la chirurgie de la cataracte ou dans le cadre de l'utilisation de robots médicaux (le robot Da Vinci par exemple).

I.1.3 Positionnement du travail de thèse dans les recherches du LaTIM

Le Laboratoire de Traitement de l'Information Médicale (LaTIM) est l'UMR 1101 de l'INSERM (Institut National de la Santé Et de la Recherche Médicale) située à Brest. Le laboratoire développe une recherche multidisciplinaire (Figure 5) dans laquelle sciences de l'information et sciences de la santé s'enrichissent mutuellement via des échanges constants entre les deux domaines. La recherche est conduite par une équipe multidisciplinaire associant des membres issus de l'Université de Bretagne Occidentale (faculté de médecine), du Centre Hospitalier Régional Universitaire (CHRU) de Brest, de l'INSERM et de Télécom Bretagne, école d'ingénieurs de l'Institut Mines-Télécom.

L'axe de recherche transversal Gestion des Données Médicales Multimodales Partagées pour l'aide à la décision (GD2MP) développe des recherches sur les bases de données médicales, à la fois pour sécuriser le partage de ces données et pour les réutiliser pour l'aide à la décision médicale. Dans cette optique, depuis plusieurs années, des études ont été réalisées dans le domaine de la recherche d'images par le contenu dans un premier temps, puis étendues à la recherche de cas cliniques contenant des données images. En parallèle, des travaux en recherche de vidéos par le contenu, pour l'aide per-opératoire, ont été initiés. Les méthodes développées ont pour champ d'application l'ophtalmologie grâce à une collaboration forte avec le service d'ophtalmologie du Centre Hospitalier Régional Universitaire (CHRU) de Brest.



a) L'Imagerie Multi-modalité quantitative pour le diagnostic et la thérapie

b) La Gestion des Données Médicales Multimodales Partagées pour l'aide à la décision

c) L'Information Physiologique pour l'Appareil Locomoteur

Figure 5. Les 3 axes du LaTIM

I.1.3.1 La Recherche d'images par le contenu (CBIR)

Les travaux menés au LaTIM en recherche d'images par le contenu ont pour objectif principal l'aide au diagnostic, notamment dans le cadre du dépistage de la rétinopathie diabétique. La rétinopathie diabétique est une complication du diabète qui atteint la rétine. Le diagnostic et le dépistage de cette maladie se fait grâce à un examen du fond de l'œil. Cet examen doit être réalisé tous les ans pour chaque patient diabétique. Un grand nombre de clichés est alors examiné chaque année par les ophtalmologistes, afin de détecter la présence et le nombre d'éventuelles lésions. Afin de faciliter et d'accélérer l'examen de ces clichés ou d'en réduire le nombre, de nombreux algorithmes d'analyse automatique des images ont et sont toujours étudiés. La littérature est très riche sur le sujet [7][8]. Ce qui caractérise la plupart des méthodes proposées, c'est la recherche de symptômes, de lésions, par segmentation, classification des pixels dans les images.

L'approche choisie par notre équipe est d'essayer de s'affranchir de ces méthodes, en utilisant des informations globales, par exemple issues de méthodes de compression d'images pour construire des signatures visuelles des images. Ainsi un premier travail a été réalisé par Ordoñez et al. autour de la compression JPEG en utilisant les coefficients de la transformée en cosinus discrète (DCT pour « Discret Cosines Transform ») et de la compression JPEG-2000 en utilisant les coefficients de la décomposition en ondelette [9]. Les travaux sur la décomposition en ondelettes ont été poursuivis par Quellec et al. [10][11][12][13]. Jai-Andaloussi et al. ont également travaillé sur la Décomposition en Modes Empiriques bidimensionnelle (BEMD pour « Bidimensional Empirical Mode Decomposition » en anglais) [14][15]. Les évaluations ont prouvé que cette approche est au moins aussi performante que la caractérisation basée sur la décomposition en ondelettes, mais coûteuse en temps de calcul. La deuxième originalité de l'approche du LaTIM est d'associer d'autres informations aux informations issues des images. Quellec et al. ont ainsi travaillé sur des méthodes d'aide à la prise de décision, fusionnant les informations provenant de l'ensemble des images du dossier, avec des informations sémantiques contextuelles telles que l'âge, le sexe ou les antécédents du patient [16]. Trois approches ont été évaluées pour fusionner ces données hétérogènes. La première s'appuie sur les arbres de décisions et leurs extensions, la seconde sur les réseaux Bayésiens et la troisième sur la théorie de Dezert-Smarandache (DSmT). Ces méthodes

ont montré leurs bonnes performances et ont prouvé qu'elles étaient capables de fournir des résultats avec un taux d'erreur identique à celui des médecins [17].

I.1.3.2 L'analyse automatique de vidéos chirurgicales

Sur la base de ces travaux, l'équipe s'intéresse depuis plusieurs années à l'exploitation des vidéos enregistrées lors de chirurgies sous contrôle vidéo telles que la chirurgie de pelage de membrane épirétinienne ou de la cataracte en ophtalmologie. Les méthodes développées ont pour objectif futur d'apporter une aide en temps réel au chirurgien, en lui proposant des exemples de situations similaires, des recommandations ou des alertes. Il faut donc être capable d'analyser en temps réel le flux vidéo enregistré pendant la chirurgie, et le comparer aux données archivées. Pour réduire et faciliter la recherche de cas similaires, nous nous appuyons sur les protocoles chirurgicaux, qui définissent les différentes étapes d'une chirurgie. Une première étape essentielle est donc d'être capable de reconnaître la tâche chirurgicale en cours d'exécution par le chirurgien. Tout comme pour les images fixes, ce sont des méthodes basées sur des méthodes de recherche par le contenu qui sont étudiées, mais cette fois sur des séquences vidéo. Dans un premier temps des méthodes de reconnaissance automatique ont été développées pour chercher à reconnaître automatiquement la tâche chirurgicale effectuée au sein d'une sous-séquence. Puis ces méthodes ont été adaptées et utilisées pour réaliser une tâche plus complexe : le séquençage automatique d'une vidéo de chirurgie complète en tâches chirurgicales. Ces méthodes sont détaillées dans le paragraphe I.2.2.

I.1.4 Conclusion

Plusieurs méthodes rapides et efficaces ont ainsi été développées au sein de l'équipe GD2MP du LaTIM autour de la réutilisation des données médicales pour l'aide à la décision, en s'appuyant plus particulièrement sur des méthodes de CBIR puis de CBVR. Pour notre objectif d'aide à la chirurgie en temps réel, plusieurs pistes ont été explorées et validées pour permettre dans un premier temps de reconnaître la tâche chirurgicale en cours d'exécution. Il s'agit d'une étape essentielle pour être à même de proposer des recommandations adaptées, de reconnaître des situations non courantes ou critiques et des améliorations restent possibles. Il existe dans la littérature plusieurs méthodes développées dans le cadre de l'analyse automatique de vidéos. Ces méthodes sont présentées dans la section suivante (paragraphe I.2).

I.2 Aide à la chirurgie en temps réel

Le LaTIM a développé des compétences reconnues dans l'utilisation du raisonnement à base de cas pour l'aide à la décision, avec la particularité de travailler avec des cas cliniques comportant des images et des vidéos, non analysées par des experts. Le sujet de cette thèse s'inscrit dans la continuité de ces travaux. L'objectif est de rechercher de nouvelles méthodes d'analyse et de caractérisation en temps réel des vidéos chirurgicales pour l'aide en per-opératoire à la chirurgie. L'état de l'art des travaux dans le domaine de l'analyse automatique de vidéos est présenté dans les parties I.2.1 et I.2.3. Puis les travaux du LaTIM déjà réalisés autour de l'analyse automatique de vidéos chirurgicales, qui ont servi de référence à nos travaux de thèse, sont détaillés dans la partie I.2.2.

I.2.1 Travaux principaux en analyse automatique de vidéos

I.2.1.1 Dans le domaine médical

L'analyse et la caractérisation de vidéos commencent à faire leur apparition dans le domaine médical, en particulier pour l'analyse de vidéos d'examens endoscopiques, mais aussi dans le cadre des chirurgies sous contrôle vidéo ou l'analyse de scènes opératoires. Ce qui explique qu'il n'y ait encore que peu de méthodes publiées sur l'analyse automatique de vidéos médicales. On constate que beaucoup de méthodes s'intéressent à l'analyse des vidéos pour leur archivage de manière efficace. Il est en effet primordial de disposer de vidéos annotées pour utiliser ces archives. Or l'annotation est une tâche complexe et surtout très coûteuse en temps de spécialiste. Les développements portent donc souvent sur des méthodes automatiques pour créer, des résumés d'examens (séquençage et sélection des événements pertinents) annoter automatiquement des vidéos (séquençage en phases ou gestes chirurgicaux) et rechercher dans les archives des cas similaires à des requêtes, pour l'aide au diagnostic.

Par ailleurs, un autre champ d'investigation se développe, qui concerne l'aide à la formation des jeunes médecins, avec notamment l'évaluation automatique des compétences. Notons enfin qu'il existe aussi des travaux sur l'analyse complète de la scène opératoire, qui intègre tous les acteurs humains et dispositifs médicaux mis en œuvre au cours d'une intervention chirurgicale. Nous présentons ci-dessous les principales approches mises en œuvre, et en donnons un récapitulatif pour le domaine médical dans le Tableau 1.

I.2.1.1.1 Résumé automatique d'examens

Plusieurs travaux, notamment dans le cas des examens endoscopiques proposent de faire un résumé automatique des examens. Stanek et al. par exemple proposent de rassembler automatiquement les différentes vidéos d'un même examen en une seule vidéo, ne contenant que les parties utiles de chaque vidéo, facilement exploitable par un médecin [18]. La méthode mise en place permet d'éliminer les images non pertinentes, c'est-à-dire prises à l'extérieur du patient. Pour cela les caractéristiques visuelles utilisées sont les couleurs de chaque image et les variations de la couleur rouge entre plusieurs paires d'images consécutives. En effet la couleur rouge est très présente lors de l'observation de nombreux tissus humains. La méthode est très rapide et de très bon

taux de reconnaissance sont obtenus, cependant le problème est un peu différent du nôtre. Il se limite à la détection de deux phases, facilement différentiables par la variation de couleurs (majorité de gris à l'extérieur du patient et majorité de rouge lors de l'examen des tissus). Cao et al. cherchent également à détecter automatiquement les événements pertinents au sein d'une vidéo d'examen endoscopique [19]. Ils s'appuient sur la détection des instruments pour détecter les passages qui correspondent à un geste diagnostique ou thérapeutique.

Pour aller plus loin que le résumé automatique d'examens, on peut également aider à l'archivage des données médicales en annotant automatiquement les vidéos contenant des informations pertinentes.

I.2.1.1.2 Annotation automatique de vidéos

Un premier type de systèmes d'annotation des vidéos d'une base de données consiste à organiser la base en classant les vidéos en différents groupes. Il s'agit alors de méthodes de classification. Twinanda et al. proposent, par exemple, une classification automatique de vidéos de chirurgies laparoscopiques basée sur une approche multi-noyaux d'un classifieur SVM (Machine à vecteurs de support ou « Support Vector Machines » en anglais) [20]. Le classifieur prend en entrée différentes caractéristiques visuelles (couleurs, descripteurs SIFT, gradients) représentées de façon compacte. Cette approche consiste à assigner à une vidéo un label. Une tâche plus complexe consiste à déterminer l'enchaînement des différentes phases chirurgicales dans une vidéo de chirurgie. Cette tâche peut être coûteuse en temps de calcul, mais dans le cas de l'aide à l'archivage des données médicales les méthodes n'ont pas de contrainte de temps réel. C'est le cas par exemple des méthodes proposées par Lalys et al. pour la reconnaissance automatique de phases [21] et d'activités [22] pour la chirurgie de la cataracte [23]. Ces méthodes s'appuient sur des signatures très complètes, basées sur l'extraction nombreuses caractéristiques visuelles (forme, couleurs, texture [21] et reconnaissance d'outils et de structures anatomiques [22]). Cependant, la construction de ces signatures est coûteuse en temps de calcul. Dans ces méthodes, deux types de modélisation du processus chirurgical ont été utilisées pour modéliser l'aspect temporel de la chirurgie. Une chirurgie moyenne construite à l'aide de l'algorithme DTW (paragraphe IV.1.2.1 page 101), qui permet de recaler temporellement deux vidéos, et les chaînes de Markov cachées (HMM) ont été utilisées pour modéliser le processus chirurgical. L'algorithme DTW nécessite de connaître l'intégralité de la vidéo pour pouvoir la recaler sur la chirurgie moyenne. Ceci n'est donc pas compatible avec une analyse en direct (ou « on line » en anglais) de la vidéo, contrairement aux chaînes de Markov cachées. Ces algorithmes d'analyse de séries temporelles (DTW avec une chirurgie moyenne et HMM) ont également été utilisés par Blum et al. [24] et Padoy et al. [25], qui travaillent sur des vidéos de cholécystectomies, une chirurgie laparoscopique visant à retirer la vésicule biliaire. L'objectif des méthodes développées est également de segmenter automatiquement en phases chirurgicales une nouvelle chirurgie. L'information de la présence des instruments est utilisée pour caractériser les vidéos, car elle est très fortement corrélée à la réalisation des différentes phases chirurgicales. Cette information est donnée par les médecins, ce qui permet de réduire considérablement les temps de calcul. L'utilisation de l'algorithme DTW, ou d'une combinaison de l'algorithme DTW avec les HMM, bien que très rapide, ne permet pas une analyse en direct de la chirurgie. En revanche, Padoy et al. proposent également une méthode uniquement basée sur les HMM, permettant une analyse en direct du flux chirurgical [25]. Forestier et al. ont également développé une méthode d'annotation automatique des vidéos en temps réel et en direct [26]. Ils cherchent ainsi à prédire quelle phase chirurgicale sera effectuée, en s'appuyant

sur un niveau de granularité plus faible (activités chirurgicales). Cette méthode propose une alternative à l'utilisation des chaînes de Markov cachées pour la modélisation du processus chirurgical en utilisant les arbres de décision. Dans ce cas, le contenu visuel de la vidéo de la chirurgie n'est pas analysé : les caractéristiques utilisées en entrée du modèle de reconnaissance étant la connaissance de l'activité en cours d'exécution.

Mackiewicz et al. utilisent quand à eux les informations de couleur pour analyser des vidéos d'examens endoscopiques par capsule [27]. Leur objectif est de déterminer automatiquement quelle structure anatomique est visitée par la capsule : l'œsophage, l'estomac, l'intestin grêle ou le côlon. Une majorité de méthodes développées dans l'analyse de vidéos d'examens endoscopiques utilise les couleurs. Fisher et Mackiewicz proposent une revue de bibliographie autour des différentes méthodes d'analyse de couleurs dans le cadre des vidéos d'examens endoscopiques par capsules [28].

Haro et al. [29], Tao et al. [30], Zappella et al. [31] quant à eux, proposent des méthodes de classification de gestes chirurgicaux en se basant sur des vidéos réalisées avec le robot Da Vinci. L'utilisation du robot permet également d'avoir accès à des informations sur les trajectoires des outils chirurgicaux dans l'espace. Cette seule information de mouvement est utilisée par Tao et al. pour la classification de gestes chirurgicaux [30]. Les méthodes développées par Zappella et al. [31] et Haro et al. [29] ont quant à elles pour objectif de classifier des gestes chirurgicaux et montrent que les méthodes basées uniquement sur la vidéo (caractéristiques visuelles 2D) donnent des résultats équivalents, voire supérieurs, aux méthodes de la littérature basées sur les trajectoires (3D) des outils chirurgicaux dans l'espace. Cela montre l'intérêt et le potentiel de l'analyse automatique de vidéos.

I.2.1.1.3 Recherche de cas similaires

L'objectif de ce type de méthodes est de simplifier la consultation des archives médicales vidéo, en s'appuyant sur les méthodes de CBVR pour rechercher et proposer automatiquement des cas similaires au cas examiné par le praticien. André et al. proposent par exemple, pour des examen vidéos-endoscopiques, une aide au diagnostic du cancer du côlon [32] par recherche de cas similaires. Pour que les informations fournies par la recherche de cas similaires soient plus facilement interprétables par les médecins, la méthode propose de transformer les signatures visuelles en contenu sémantique. Cette approche est particulièrement intéressante par sa façon originale d'apporter l'information aux médecins et de lier les informations visuelles aux informations sémantiques. En revanche l'approche s'appuie sur des méthodes de recherche d'images par le contenu en s'appuyant sur une méthode classique de CBIR : la création de sacs de mots visuels pour caractériser chacune des images. L'aspect dynamique de la vidéo n'est pas pris en compte, et c'est pourtant un élément qui semble important pour caractériser une chirurgie.

I.2.1.1.4 Evaluation des compétences des chirurgiens

L'analyse automatique de vidéos trouve également sa place dans l'aide à la formation des jeunes médecins, par exemple dans le cas de l'utilisation de simulateurs. Ainsi Oropesa et al. proposent une méthode d'évaluation des compétences basée sur une étude du mouvement des instruments chirurgicaux au sein d'un simulateur de chirurgies laparoscopiques [33]. La détection et le suivi des instruments se font par analyse vidéo. Différents critères sont ensuite évalués, tels que

la vitesse et l'accélération des instruments, l'espace utilisé pour la réalisation des gestes chirurgicaux, la durée des phases de transition, etc. Leong et al. ont également travaillé à l'étude de la qualité du mouvement et des trajectoires des instruments dans le cas des chirurgies laparoscopiques en s'appuyant sur les chaînes de Markov cachées comme modèle statistique [34]. Enfin, Suzuki et al. ont développé un logiciel d'analyse du mouvement pour l'évaluation des compétences chirurgicales en chirurgie endoscopique [35]. Ils s'appuient également sur différents critères tels que le temps d'exécution des tâches chirurgicales ou la longueur des trajets des instruments chirurgicaux pour évaluer le niveau de compétence.

Reiley et al. cherchent à reconnaître automatiquement le niveau de compétence du chirurgien (débutant, intermédiaire ou expert) [36][37]. Reiley et al. proposent une revue de la littérature de ce champ d'études [38]. Ce domaine d'étude est un peu différent du nôtre mais cela requiert également d'analyser automatiquement les gestes chirurgicaux. Reiley et al. ont par exemple étudié deux niveaux de description différents (tâches et gestes réalisés par le robot Da Vinci) en les modélisant par des chaînes de Markov cachées (HMMs) [36]. Les résultats montrent que le niveau de description le plus fin (en gestes chirurgicaux) permet de mieux identifier ce qui a été bien ou moins bien réalisé par les chirurgiens. Tao et al. ont également travaillé à l'évaluation automatique des compétences à partir de données fournies par le robot Da Vinci [39]. Un dictionnaire des différents gestes chirurgicaux a été construit à partir des données de mouvement enregistrées par différents chirurgiens de niveaux de compétence variés. Les transitions entre les différents gestes ont également été modélisées par une chaîne de Markov cachée (HMM). La combinaison entre le dictionnaire et la grammaire associée (HMM) a montré des résultats intéressants.

I.2.1.1.5 L'analyse de scènes opératoires

Plus largement que l'étude du flux vidéo visualisé par le médecin ou les aides opératoires, il existe des travaux d'analyse de flux vidéo représentant une scène opératoire complète. Garraud et al. ont par exemple travaillé sur l'analyse de scènes opératoires et ont développé une suite logicielle basée sur une ontologie permettant de modéliser tout le processus chirurgical [40]. Cela prend en compte tous les événements qui ont lieu dans la salle opératoire et ne s'appuie donc pas uniquement sur la vidéo, mais aussi sur les signaux vitaux du patient, etc. Bathia et al. eux, cherchent à déterminer automatiquement l'occupation des salles opératoires en s'appuyant sur un modèle de Markov cachée (HMM) [41]. Ce travail est un peu différent du nôtre car il n'analyse pas la chirurgie elle-même, mais il est néanmoins très pertinent dans notre cas car il réalise une analyse en temps réel et en direct (c'est-à-dire tout au long de l'acquisition de la vidéo).

I.2.1.1.6 Conclusion

Peu de travaux sont réalisés actuellement dans le domaine médical, qui soient capable d'analyser et de caractériser les données en temps réel, durant l'acquisition de la vidéo. Or, l'aspect temps réel est essentiel dans notre approche. Apporter une aide au cours de la chirurgie impose d'analyser la vidéo pendant son acquisition. On parle dans ce cas de méthodes « *en direct* » (« on line » en anglais). Le corollaire est que les algorithmes doivent être extrêmement *rapides*. Ces deux aspects sont détaillés dans le paragraphe I.2.3. Les principales méthodes sont présentées dans le Tableau 1.

Tableau 1. Tableau récapitulatif des différentes méthodes proposées en analyse automatique de vidéos dans le domaine médical

Auteurs	Journal/conf	année	on line	off line	Rapide	Objectifs
Leong et al.	MICCAI	2006		✓		Evaluation des compétences
Cao et al.	IEEE TBE	2007		✓		Segmentation spatiotemporelle
Bathia et al.	IAAI	2007	✓		✓	Analyse de scènes opératoires
Mackiewicz et al.	IEEE TMI	2008		✓		Séquençage automatique en phases pertinentes -Archivage
Reiley et al.	MICCAI	2009		✓		Evaluation de compétences
Lalys et al.	IEEE TBE	2012		✓		Séquençage automatique en phases chirurgicale - Archivage
Andre et al.	IEEE TMI	2012		✓		Recherches de cas similaires
Padoy et al.	Elsevier MIA	2012	✓	✓	✓	Séquençage automatique en phases chirurgicales
Tao et al.	IPCAI	2012		✓		Classification de gestes chirurgicaux - évaluation des compétences
Droueche et al.	IEEE EMBC	2012		✓		Classification de tâches chirurgicales
Staneck et al.	Elsevier CMPB	2012			✓	Séquençage automatique en phases pertinentes - Archivage
Haro et al.	MICCAI	2012		✓		Classification de gestes chirurgicaux
Lalys et al.	CARS	2013		✓		Séquençage automatique en phases chirurgicales - Archivage
Zappella et al.	Elsevier MIA	2013		✓		Classification de gestes chirurgicaux
Oropesa et al.	Surg Endosc	2013		✓		Evaluation des compétences
Tao et al.	MICCAI	2013		✓		Séquençage automatique en geste chirurgicaux

Garraud et al.	Surgetica	2014	✓	✓	✓	Analyse de scènes opératoires
Quellec et al.	MIA	2014			✓	Classification de tâches chirurgicales
Quellec et al.	TMI	2014	✓		✓	Séquençage automatique en tâches chirurgicales – aide à la chirurgie
Suzuki et al.	Surg Endosc	2014		✓		Evaluation de compétences
Forestier et al.	CARS	2015	✓		✓	Séquençage automatique en phases pertinentes
Quellec et al.	MIA	2015	✓		✓	Classification de tâches chirurgicales
Twinanda et al.	CARS	2015		✓		Classification de vidéos chirurgicales

Il existe en revanche de nombreuses méthodes permettant de réaliser ce type d'analyse temps réel et « en direct » dans d'autres domaines, tels que la vidéosurveillance.

I.2.1.2 Analyse de séquences vidéos dans d'autres domaines

La télésurveillance fait largement appel aux données vidéo. Ces données sont présentes en grande quantité et difficilement exploitables manuellement. Il est de plus nécessaire de les analyser en temps réel pour reconnaître et anticiper les situations anormales. Cela est difficilement réalisable et requiert un grand nombre de personnes pour analyser en temps réel différentes sources vidéo. De nombreuses méthodes ont donc été développées pour l'analyse automatique de ce type de vidéos, notamment pour la détection d'événements atypiques, situations qui nous intéressent pour le suivi des actes chirurgicaux.

Un premier type d'analyse concerne la reconnaissance automatique de visages [42]. Il s'agit généralement de méthodes de détection, de suivi et de reconnaissance de formes, qui sont un peu plus éloignées de ce que l'on souhaite faire. Néanmoins ces méthodes peuvent être une bonne source d'inspiration si nous souhaitons suivre et reconnaître les instruments chirurgicaux en cours d'utilisation. De nombreuses méthodes s'attellent également à l'étude automatique du trafic routier ou des flux de piétons. Castaton et al. s'appuient par exemple sur l'extraction de caractéristiques utilisées dans une table de hachage pour effectuer une recherche extrêmement rapide des cas similaires [43]. Cette méthode très rapide se montre efficace pour la détection d'abandon d'objet, le comptage, ou la reconnaissance de schémas de mouvements typiques. Piciarelli et al. réalisent, quand à eux, un partitionnement de trajectoires structurées en arbres pour réaliser une détection automatique des trajectoires atypiques de véhicules [44]. Acevedo-

Rodríguez et al. travaillent également à l'étude des trajectoires, mais des piétons [45]. Les trajectoires sont regroupées en différents groupes de trajectoires typiques. En effet, en détectant les mouvements et les trajectoires typiques, il est plus aisé de repérer par la suite les évènements anormaux, c'est-à-dire s'éloignant de la réalité connue. Ainsi [46] s'appuient sur des méthodes probabilistes bayésiennes pour analyser automatiquement les comportements de piétons ou de véhicules en temps réel. La méthode permet d'apprendre des schémas classiques de comportements et est ainsi capable de détecter des événements atypiques. Ces méthodes sont intéressantes par leur capacité à distinguer des événements atypiques par rapport à la normalité précédemment apprise. Cependant, les vidéos de surveillance sont relativement différentes des vidéos de chirurgies : elles filment généralement une scène dont le second plan est fixe et dans ce type de vidéos, les trajectoires attendues sont facilement identifiables.

I.2.1.3 Conclusion

Pour l'annotation automatique de vidéos médicales ou l'étude de vidéos de télésurveillance, les méthodes de CBVR semblent fournir une réponse intéressante. Elles permettent de présenter des cas similaires et donc d'apporter une information sur une vidéo ou un segment de vidéo. Une majorité de méthodes mises en place utilisent également des méthodes prédictives de fouille de données ou d'analyse de séries temporelles telles que les chaînes de Markov cachées, particulièrement adaptées à l'étude de processus temporels. Cela permet de prendre en compte l'aspect dynamique apporté par la vidéo, mais aussi d'apporter une information contextuelle. Il se dégage également que l'utilisation du mouvement comme caractéristique visuelle semble particulièrement adaptée. Il apparaît également qu'il est difficile de comparer les méthodes développées pour le domaine médical car il n'y a actuellement pas de bases de données communes : les méthodes sont évaluées sur des bases de données très diverses, non publiques en général. Les différentes bases de données existantes dans la littérature ainsi que leurs méthodes de description sont détaillées dans le paragraphe II.1.2 page 48.

I.2.2 Les travaux du LaTIM dans l'analyse automatique de vidéos chirurgicales

Les travaux du LaTIM dans le domaine de l'analyse automatique de vidéos chirurgicales s'appliquent actuellement à la chirurgie de la cataracte. Dans un premier temps des méthodes de reconnaissance automatique ont été développées pour chercher à reconnaître automatiquement une tâche chirurgicale effectuée au sein d'une sous-séquence. Ces méthodes s'appuient sur une recherche de cas similaires. Puis ces méthodes ont été adaptées et utilisées pour annoter automatiquement des vidéos de chirurgies complètes. Les vidéos sont alors séquencées automatiquement en tâches chirurgicales, et cela « en direct ».

I.2.2.1 Reconnaissance automatique de tâches chirurgicales

Dans un premier temps, pour évaluer les méthodes de reconnaissance automatique de tâches et particulièrement les étapes clés de la CBVR (caractérisation et mesure de similitude), le pro-

blème a été simplifié. Plutôt que de travailler directement à l'analyse de vidéos de chirurgies complètes, la problématique a été ramenée à la classification de séquences vidéo dont le principe est présenté dans la Figure 6. Les vidéos ont été découpées en séquences où chacune des séquences représente une tâche chirurgicale. Plusieurs méthodes ont alors été développées au sein de l'équipe pour chercher à reconnaître automatiquement la tâche chirurgicale réalisée dans la séquence, en s'appuyant sur des méthodes de CBVR.

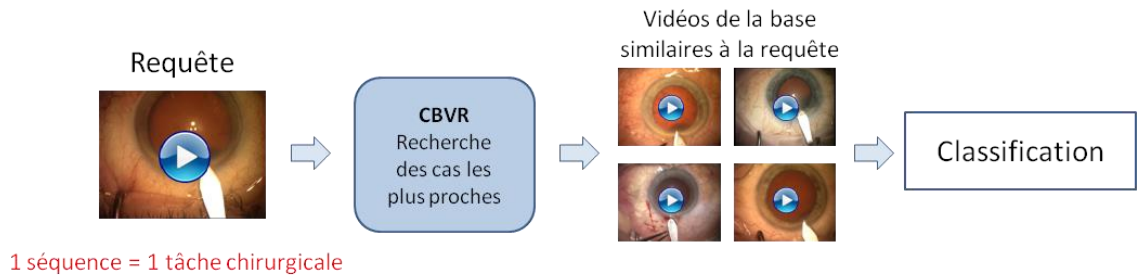


Figure 6. Principe des méthodes de reconnaissance automatique de tâches

Une première méthode développée par Z. Droueche et al. s'inspire des méthodes précédemment développées par l'équipe en CBIR. Elle utilise les informations issues de la compression MPEG (« Moving Picture Experts Group » en anglais) pour construire les caractéristiques visuelles de chaque séquence vidéo [47][48]. Le mouvement des blocs de l'image est obtenu grâce à la compression et utilisé pour construire et suivre des régions de mouvement homogène (Figure 7).

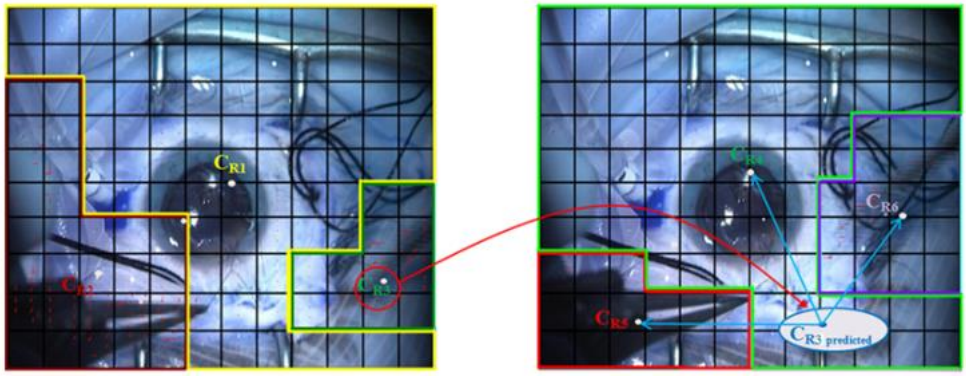
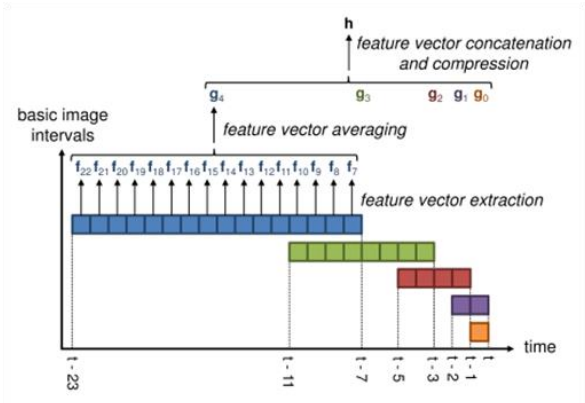


Figure 7. Exemple de suivi de région entre deux images consécutives [48]

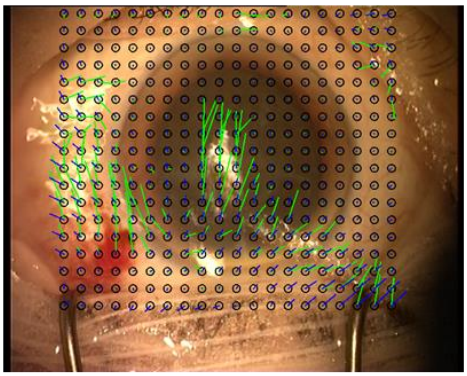
Le suivi de ces régions (position et vitesse des centres, direction dominante) permet de construire les signatures visuelles des séquences vidéo. Une évolution de l'algorithme classique DTW (« Dynamique Time Warping ») a ensuite été utilisée pour mesurer la similitude entre deux vidéos. Il s'agit de l'algorithme EFDTW (« Extented Fast Dynamic Time Warping ») qui est une combinaison entre l'algorithme FDTW (« Fast Dynamique Time Warping ») avec la mesure de distance EMD (« Earth Mover's Distance »). Les performances de recherche ont été évaluées en termes de précision. La précision représente le nombre de cas pertinents retrouvés parmi les cas proposés par l'outil de recherche pour une requête donnée. Les résultats obtenus avec les deux bases de vidéos de chirurgies oculaires (pelage de membrane et cataracte) sont encourageants, avec une précision

moyenne pour une fenêtre de 5 cas de 79,0 % pour la base de pelage de membrane et de 72,7 % pour la base de cataracte. Bien que relativement rapide, cette méthode reste coûteuse en temps de calcul et n'autorise donc pas le temps réel.

Plusieurs méthodes adaptées au temps réel (et l'analyse « en direct ») ont également été développées par G. Quellec et al. Dans une première méthode [49], l'idée est de découper automatiquement un geste chirurgical en mouvements élémentaires pour faciliter sa reconnaissance. Dans ce système, de courtes sous-séquences, extraites du flux vidéo, sont caractérisées puis comparées à des sous-séquences archivées. Les sous-séquences sont caractérisées par des vecteurs de taille fixe, construits de façon à ce que les caractéristiques soient inchangées en fonction des variations de durée et de vitesse d'exécution au sein des tâches chirurgicales (Figure 8, à gauche). Cela permet une recherche rapide des plus proches voisins dans la base de données. Ce système est très rapide et fournit de bons taux de reconnaissance. Une seconde méthode consiste à approcher le champ des vecteurs de déplacement au cours d'une courte séquence vidéo par un polynôme spatiotemporel, dont les paramètres servent de signature (Figure 8, à droite) [50]. Pour chaque tâche chirurgicale visée, un apprentissage est effectué pour identifier quels polynômes spatiotemporels sont généralement extraits quand cette tâche est réalisée dans la séquence vidéo. Ces polynômes spatiotemporels sont ensuite recherchés dans la séquence vidéo requête pour identifier la tâche chirurgicale réalisée. Cette méthode évaluée sur la base de chirurgies de la cataracte donne de très bons résultats avec une aire moyenne sous la courbe ROC (« Receiver Operating Characteristic » en anglais) de 0,851 [50].



G. Quellec, MIA 2014



G. Quellec, IEEE TMI 2015

Figure 8. A gauche, extraction des vecteurs de caractéristiques des sous-séquences, permettant de découper automatiquement un geste chirurgical en mouvements élémentaires [49]; A droite, en bleu, le champ de mouvements approché par des polynômes spatiotemporels, en vert les champs de mouvements entre deux images consécutives mesurés par l'algorithme de Farnebäck [50]

Plusieurs méthodes très rapides ont ainsi été évaluées et validées pour reconnaître automatiquement la tâche chirurgicale représentée dans une séquence vidéo. Ces méthodes s'appuient sur des systèmes de CBVR et plusieurs méthodes de caractérisation et de comparaisons ont ainsi été évaluées pour chercher les cas similaires dans la base de vidéos archivées. Cependant un autre défi consiste à reconnaître en temps réel la tâche chirurgicale effectuée au sein d'une vidéo de chirurgie en cours d'exécution. Cela nécessite de mettre en place des méthodes de séquençage automatique des vidéos en tâches chirurgicales. Des méthodes ont déjà été implémentées au sein du LaTIM dans ce domaine.

I.2.2.2 Séquençage automatique de vidéos de chirurgies

Pour analyser automatiquement une vidéo de chirurgie en cours d'exécution, il est nécessaire de détecter automatiquement le démarrage et la fin de la réalisation de chacune des tâches chirurgicales. Une méthode temps réel de segmentation automatique en tâches chirurgicales a été proposée par G. Quellec et al. [51]. Cette méthode s'appuie sur le fait qu'il existe généralement un délai de transition entre deux tâches chirurgicales, durant lequel il ne se passe rien de pertinent dans le champ de vue de la caméra. Ce délai vient du fait que le chirurgien change d'instruments entre deux tâches chirurgicales. La méthode proposée travaille donc dans un premier temps à la détection de ces transitions, en s'appuyant sur une méthode de recherche de plus proches voisins. La chirurgie est alors segmentée temporellement en phases d'action et en phases de transition. A chaque fois qu'une phase de transition est détectée, la phase d'action qui la précède est classifiée (Figure 9, à gauche). Les champs markoviens conditionnels (CRF pour « Conditional Random Fields » en anglais) ont été utilisés pour cette étape de catégorisation. Les caractéristiques de mouvement entre l'image en cours et la précédente, les caractéristiques de couleur et de texture ainsi que la durée de la phase de transition sont utilisés pour construire les signatures visuelles des segments (Figure 9, à droite).

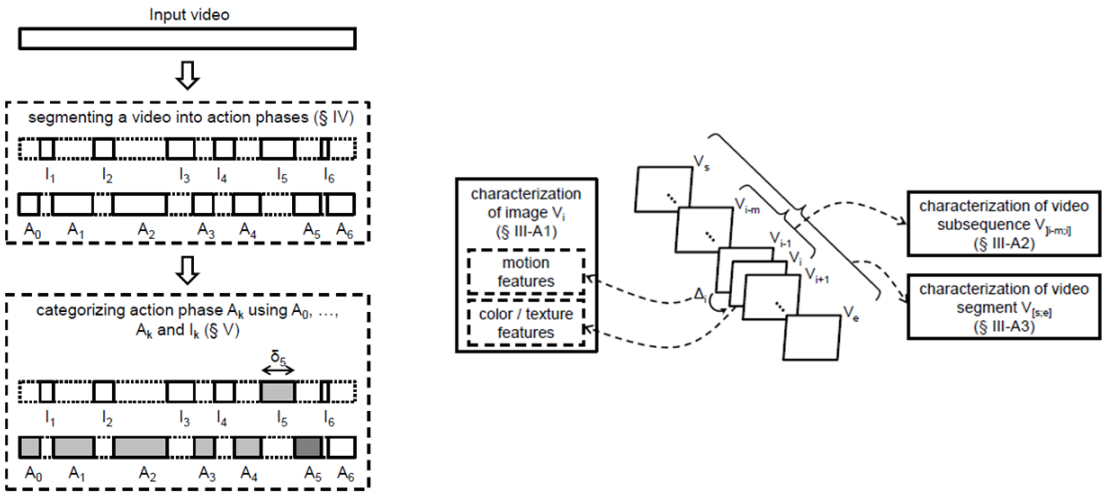


Figure 9. A gauche, le principe de la méthode de segmentation et de catégorisation des sous-séquences proposée par G. Quellec et al. [51], à droite, la méthode de caractérisation associée

Cette méthode, évaluée sur la base de vidéos de chirurgies de la cataracte montre de bonnes performances, avec une aire moyenne sous la courbe ROC de 0,832. Sa principale limite est que plusieurs tâches chirurgicales peuvent avoir lieu pendant une phase d'action. C'est le cas si la phase de transition n'est pas détectée entre deux tâches chirurgicales ou si le chirurgien a changé d'outil pour une main, tout en continuant une action avec l'autre main.

Ces méthodes mises en place par le LaTIM ont servi de base à mon travail, dont l'objectif est également de segmenter automatiquement « en direct » les vidéos de chirurgies de la cataracte. Mes objectifs se sont tournés vers l'amélioration des méthodes de reconnaissance, et la mise en place d'une analyse plus fine et plus complète de la chirurgie.

I.2.3 L'analyse automatique de vidéos en temps réel

Comme cela a déjà été évoqué précédemment, l'aspect temps réel est un point essentiel de notre approche, indispensable pour apporter une aide tout au long de la chirurgie. Cela se traduit d'abord par la nécessité d'avoir de« algorithmes rapides, afin d'analyser la vidéo pendant son acquisition et fournir des informations de manière instantanée. Nous sommes dans le cadre de méthodes « *en direct* » (« on line » en anglais). Ces méthodes présentent de plus une contrainte, par rapport aux méthodes d'analyse en différé ("off line") : on ne peut qu'utiliser l'information passée ou présente. Ces deux aspects sont présentés dans les deux paragraphes suivants (I.2.3.1 et I.2.3.2).

I.2.3.1 Algorithmes rapides

Il y a deux facteurs limitants concernant les temps de calcul. Le premier concerne l'extraction des caractéristiques et le second la comparaison avec l'ensemble des vidéos ou sous-séquences de la base. Il n'est donc pas possible d'utiliser des méthodes d'extraction de caractéristiques trop complexes. Des signatures visuelles simples telles que des histogrammes de couleurs ou de mouvements sont généralement utilisés [51][41]. Dans certains cas, les auteurs s'affranchissent de l'étape d'extraction de caractéristiques visuelles, en utilisant comme signatures l'information de présence des instruments dans la scène opératoire [25][26]. Cette information est supposée être fournie dans le futur par des radio-étiquettes (puces RFID, « radio frequency identification » en anglais) par exemple. Enfin, il existe des algorithmes de recherche par le contenu très rapides, accélérés notamment par l'utilisation de tables de hachage [52][53][54]. Les tables de hachage accélèrent la recherche de cas similaires en associant une clé à chaque événement. Il s'agit d'une méthode de structuration des données qui pourra être à l'avenir une technique d'accélération de nos méthodes. Droueche et al. ont également proposé l'algorithme EFDTW qui permet une mesure de distance rapide entre deux vidéos. Cependant il s'agit d'une combinaison entre l'algorithme FDTW avec la mesure de distance EMD, il n'est donc pas compatible avec l'analyse « en direct » d'une vidéo.

I.2.3.2 Algorithmes d'analyse « en direct »

Dans le cas des algorithmes d'analyse « en direct », l'analyse doit se faire sans connaître la fin de la chirurgie. Or la majorité des méthodes d'analyse utilisent l'information contextuelle, en ne s'appuyant pas uniquement sur l'information extraite au temps courant, mais également les informations suivantes et précédentes. Les méthodes choisies pour notre problème peuvent donc s'appuyer sur les informations précédentes, mais pas sur les informations suivantes, puisque ces informations ne sont pas encore disponibles au temps courant. Les méthodes basées sur la mesure de distance DTW [21][47][25], par exemple, ne peuvent pas être utilisées car elles nécessitent de connaître l'intégralité de la vidéo requête pour la recaler temporellement avec une vidéo de la base de données ou une vidéo moyenne. Piciarelli et al. proposent une alternative intéressante à cette mesure de distance, en proposant une mesure de distance à la volée, c'est-à-dire qui est mise à jour à chaque nouvelle image acquise [44]. Il est également possible de considérer la vidéo

de chirurgie complète comme une succession de sous-séquences, qui peuvent éventuellement se chevaucher. C'est l'approche adoptée par Quellec et al. [51]. Cette technique a l'avantage de permettre d'appliquer sur ces sous-séquences vidéo les méthodes de reconnaissance automatique issues de la CBVR par exemple. Enfin, l'utilisation de modèles markoviens tels que les chaînes de Markov cachées [25] ou le MCTM (« Markov Clustering Topic Model ») de Hospedales et al. [46] est particulièrement adaptée à l'analyse « en direct ». Ces modèles permettent notamment de déduire la phase chirurgicale en cours ou à venir. Les chaînes de Markov cachées sont les modèles statistiques les plus utilisés car totalement adaptés à l'étude de séries temporelles. Elles permettent de modéliser les transitions entre les différentes phases chirurgicales. Néanmoins elles présentent certaines limites, notamment en cas de faibles nombres d'exemples dans la base de données. Des alternatives aux chaînes de Markov également adaptées à l'analyse « en direct» sont donc proposées dans la littérature, telles que les CRFs utilisés par [30] et [51], ou les arbres de décisions [26].

I.2.3.3 Conclusion

Ces différentes considérations seront donc à prendre en compte pour les différents scénarios possibles que nous pouvons envisager. Les méthodes choisies et développées doivent être rapides et ne peuvent pas s'appuyer sur des méthodes qui nécessitent de connaître l'intégralité de la vidéo. Il n'est donc possible d'utiliser que les informations du temps présent et du passé pour prendre une décision, les informations enregistrées dans le futur n'étant pas encore disponibles.

I.2.4 Scénario envisagé

L'étude des différentes méthodes de la littérature, ainsi que les différentes méthodes déjà développées au sein du LaTIM, permettent d'envisager la mise en place d'un scénario d'aide chirurgicale en temps réel. Il s'appuiera essentiellement sur des vidéos chirurgicales préalablement archivées et interprétées pour contrôler le déroulement des actes chirurgicaux, et générer des préconisations et des alertes. Ces vidéos archivées et interprétées vont nous permettre d'effectuer un raisonnement à partir de cas en cherchant des cas similaires. Le principal défi de ce travail est de développer des méthodes d'indexation et de recherche de vidéos capables de fournir des résultats en temps réel, en direct. Les méthodes proposées sont évaluées sur des vidéos de chirurgies de la cataracte collectées au sein du service d'ophtalmologie du CHRU de Brest. Ces méthodes doivent donc pouvoir gérer des vidéos de chirurgies effectuées par différents chirurgiens, plus ou moins expérimentés et enregistrées par des systèmes d'acquisition variables. Les vidéos collectées sont donc hétérogènes en durée et en qualité et les techniques chirurgicales peuvent varier pour une même étape de la chirurgie. Notre scénario repose sur deux étapes : la reconnaissance du geste chirurgical, puis la détection d'évènements anormaux et la génération d'alertes.

I.2.4.1 Reconnaissance du geste chirurgical

Il est tout d'abord nécessaire de reconnaître, à chaque instant, l'étape actuelle de la chirurgie. Le scénario envisagé pour arriver à cela est présenté dans la Figure 11. L'équipe GD2MP a acquis de solides compétences dans la *recherche de cas similaires par le contenu*, il semble donc naturel

de s'inspirer de ces méthodes pour chercher à reconnaître automatiquement le geste chirurgical effectué. Cette étape consiste à caractériser chaque image ou chaque segment de la vidéo par une signature visuelle. Grâce à une mesure de similitude adaptée, les cas les plus proches sont trouvés au sein de la base de données précédemment archivée. Ainsi une probabilité d'appartenance à une tâche peut être associée à chaque image ou chaque segment. Une base de séquences pré-segmentées manuellement, où une séquence représente une tâche chirurgicale, a été construite pour évaluer dans un premier temps nos méthodes de catégorisation. Cette partie de mise en place et d'évaluation des méthodes de caractérisation et de mesure de similitude fera l'objet du Chapitre III. Une idée est ensuite de considérer la vidéo de chirurgie complète comme une succession d'images ou de sous-séquences indépendantes et d'appliquer les méthodes de CBIR ou de CBVR présentées précédemment pour chercher les exemples les plus proches et déterminer ainsi à quelle tâche elles appartiennent Cette méthode ne permet pas en revanche de prendre en compte l'information contextuelle telle que le temps écoulé depuis le début de la chirurgie, les tâches chirurgicales qui ont été précédemment détectées, etc… Or ces informations sont particulièrement pertinentes. Il est alors intéressant de s'appuyer sur des algorithmes prédictifs de fouille de données (Figure 10). Ces méthodes s'appuient sur des modèles statistiques appris à partir des données de la base d'archives.

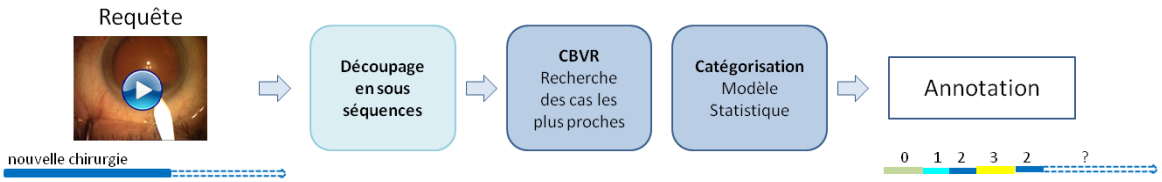


Figure 10. Principe des méthodes de séquençage automatique des vidéos de chirurgies en tâches chirurgicales développées par l'équipe GD2MP

Dans notre cas il est particulièrement intéressant de modéliser le déroulement temporel de la chirurgie, en modélisant les probabilités de transition d'une tâche chirurgicale vers une autre. De plus, comme nous souhaitons apporter les informations les plus pertinentes possible aux chirurgiens, il est important de décrire la chirurgie de la façon la plus complète possible. Ainsi, en nous inspirant des travaux de Lalys [23] et de Forestier et al. [26], nous avons travaillé à une description de la chirurgie à plusieurs niveaux de granularité (II.3). En construisant notre modèle statistique de façon à intégrer ces différents niveaux de granularité et à les faire communiquer entre eux, la reconnaissance sera d'autant plus pertinente. Cette partie de reconnaissance et de séquençage de la chirurgie à plusieurs niveaux de description fait l'objet du Chapitre IV.

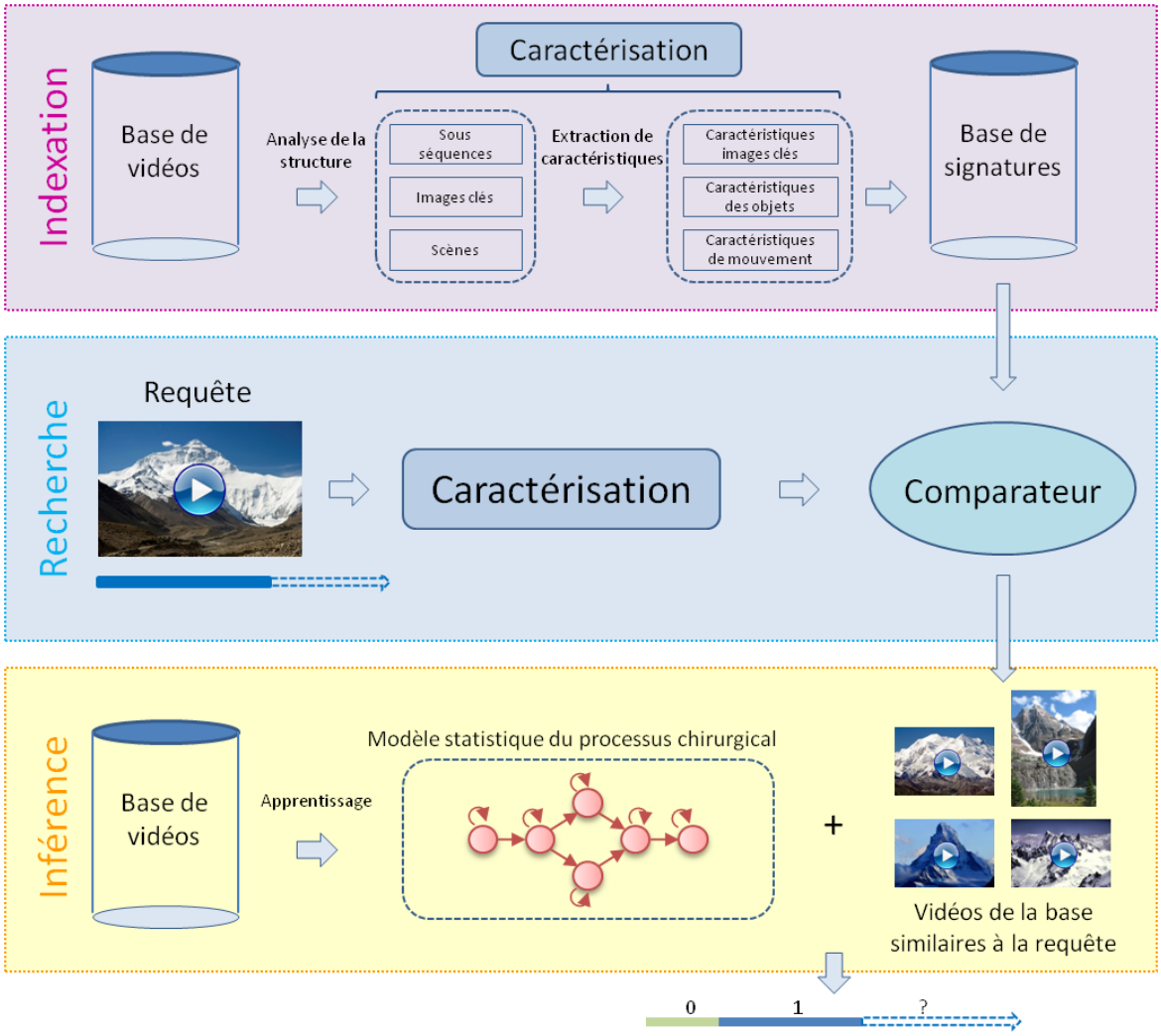


Figure 11. Scénario envisagé pour la reconnaissance du geste chirurgical

I.2.4.2 Détection d'événements anormaux et génération d'alertes

Il est aussi nécessaire d'être capable de détecter des événements (ou des séquences d'événements) annonciateurs d'une complication. Ces événements seront recherchés, par des algorithmes de fouille de données, dans des archives de vidéos chirurgicales renseignées. Ces événements pourront ainsi être détectés par des méthodes similaires à celles utilisées pour la reconnaissance du geste chirurgical. Deux états peuvent par exemple être appris : « normal » et « anormal ». On peut également imaginer plusieurs classes d'événements « normaux » et « anormaux ». A partir des actions effectuées par des médecins expérimentés dans des situations similaires, on cherchera à définir des préconisations, voire déclencher des alertes et proposer des actions. Notons qu'il s'agit là d'une perspective, l'objectif de cette thèse étant principalement porté sur une reconnaissance efficace du geste chirurgical.

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

I.3 Discussion - Conclusion

Le LaTIM a développé de nombreuses méthodes autour de la gestion et de l'utilisation des données médicales pour l'aide au diagnostic. Il est en effet extrêmement pertinent, au vu des quantités importantes de données numériques archivées et des ressources disponibles pour les analyser, de réutiliser ces données pour la consultation de cas similaires, l'extraction des connaissances, la prédiction ou l'aide au diagnostic. Les travaux développés s'appuient sur des méthodes de fouille de données et de recherche par le contenu de cas similaires dans une base de données connues et archivées. Les recherches ont essentiellement porté sur l'analyse automatique d'images et depuis peu, les méthodes ont été adaptées pour l'analyse vidéo. Nous cherchons à reconnaître automatiquement, en temps réel, le geste ou la tâche chirurgicale qui est en train d'être réalisée par le chirurgien. C'est l'objet de cette thèse. Les méthodes développées devront permettre ensuite d'alerter le chirurgien sur les déroulements opératoires à risques, et lui fournir des recommandations en temps réel sur des conduites à tenir reconnues.

Les méthodes de la littérature autour de l'analyse automatique de vidéos médicales sont rarement compatibles avec l'analyse « en direct » de la chirurgie, c'est-à-dire pendant que le médecin pratique sa chirurgie ou son examen. Ce point est un élément clé de cette thèse, et les méthodes développées devront être rapides et être capables d'analyser la chirurgie sans en connaître la fin.

Le scénario envisagé pour répondre à cette problématique s'inspire des méthodes existantes de la littérature. Il consiste à s'appuyer sur des méthodes de recherche par le contenu des cas les plus proches dans la base de données, combinées avec un modèle statistique du déroulement de la chirurgie qui apportera une information contextuelle.

Nous présenterons ensuite le domaine d'application de nos méthodes et la base de données collectée et annotée par le LaTIM en collaboration avec le service d'ophtalmologie du CHRU de Brest. Nous étudierons ensuite une méthode de recherche de cas similaires pour la reconnaissance automatique des tâches chirurgicales. Nous chercherons alors à reconnaître automatiquement la tâche réalisée dans la séquence présentée en requête. Nous étudierons enfin la construction et l'inférence de deux modèles probabilistes du déroulement de la chirurgie de la cataracte. On cherchera ici à segmenter temporellement une vidéo de chirurgie complète, à différents niveaux de description (activités, étapes et phases).

Chapitre II. Bases de données

CHAPITRE II. BASES DE DONNEES 46

II.1 Les bases de données et leurs description dans la littérature 47
II.1.1 Granularité 47
II.1.2 Les différentes descriptions des bases de données de vidéos médicales 48
II.1.2.1 Travaux de l'équipe VisAGeS 49
II.1.2.2 Autres travaux 49

II.2 La base de données du LaTIM 53
II.2.1 Application clinique : la chirurgie de la cataracte 53
II.2.2 Les données 54
II.2.2.1 Description de la chirurgie 54
II.2.2.2 Les bases de données annotées 56

II.3 Nouvelle description multi-échelles de la chirurgie 57
II.3.1 Phases 57
II.3.2 Etapes 58
II.3.3 Activités 59

II.4 Diagrammes de transition obtenus 60
II.4.1 Phases 60
II.4.2 Tâches 61
II.4.3 Etapes 61
II.4.4 Activités 62

II.5 Discussion - Conclusion 64

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

Pour évaluer les différentes méthodes présentées dans le chapitre précédent, les différents auteurs ont dû constituer des bases de données. Cependant, aucune de ces équipes n'a décidé de rendre sa base de données publique. Il n'existe donc pas de base de données commune de vidéos médicales. Chaque équipe travaille donc avec sa propre base de données et son propre système de description du processus chirurgical. Il existe donc différentes bases de données et descriptions dans la littérature. Le LaTIM a également créé sa propre base de données, contenant des vidéos de chirurgie de la cataracte, en collaboration avec le service d'ophtalmologie du CHRU de Brest. Dans un premier temps, cette base a été annotée selon une description en 9 tâches chirurgicales principales. Il est cependant nécessaire, pour apporter une aide précise aux chirurgiens, d'analyser la chirurgie avec plus de précision. C'est pourquoi avec un interne en chirurgie, nous avons réfléchi à une nouvelle description de la chirurgie de la cataracte. Les différentes bases de données médicales présentes dans la littérature sont présentées dans le paragraphe suivant (II.1). La base de données collectée par le LaTIM est présentée dans le paragraphe II.2. Puis notre nouvelle description de la chirurgie ainsi que les diagrammes de transition obtenus sont présentés dans les paragraphes II.3 et II.4.

II.1 Les bases de données et leurs description dans la littérature

Pour utiliser les données collectée afin d'analyser ensuite automatiquement une nouvelle vidéo, il est nécessaire d'annoter ces données. Dans notre cas, nous cherchons à reconnaître automatiquement les différents gestes et étapes de la chirurgie. C'est pourquoi, nous devons annoter les vidéos selon une description bien définie de la chirurgie. Il existe, dans la littérature, différentes bases de données mais aussi différentes manières de décrire une chirurgie. Il est possible de décrire la chirurgie à d'autres niveaux de précision. On parle alors de niveaux de granularité.

II.1.1 Granularité

Les différents niveaux de granularité que l'on peut trouver dans la littérature [55] sont présentés dans la Figure 12.

Figure 12. Différents niveaux de granularités que l'on peut trouver dans la littérature

Le niveau de description le plus bas niveau correspond aux **caractéristiques visuelles** extraites de la vidéo. Elles ne contiennent pas d'informations sémantiques et représentent uniquement ce qui se passe visuellement dans la vidéo. Elles sont induites par les actions ou les gestes réalisés avec les instruments chirurgicaux. Une **action** est associée à un verbe, et elle peut être vue comme l'application d'un **geste** à la réalisation de quelque chose. Le terme de « geste chirurgical » est

utilisé par [30][31][29]. Leurs gestes chirurgicaux sont définis de la manière suivante : « positionner l'aiguille », « orienter l'aiguille », « passer l'aguille à travers le tissue », etc… Tous ces gestes sont effectivement définis avec un verbe d'action. Une **activité** se définit quant à elle, d'après Lalys et al. [23], comme un triplet < verbe d'action – outil chirurgical – structure anatomique >. Les termes « étapes », « tâche », ou « phases » sont également souvent trouvés dans la littérature. Une **tâche** correspond à un travail qui doit être réalisé avec un objectif précis. Elle répond donc à la réalisation d'un objectif chirurgical, tel que réaliser une incision ou faire une suture. Une même tâche peut être réalisée plusieurs fois dans une même chirurgie. Dans le cas de la chirurgie de la cataracte par exemple, la tâche « Injection du visqueux » est réalisée plusieurs fois, pour redonner du tonus à l'œil quand cela est nécessaire. Il est également parfois nécessaire de revenir à la tâche « Incision » lors de la mise en place de l'implant, si l'incision n'est pas suffisamment grande. Nous pouvons définir une **étape** comme un épisode d'une progression, qui n'aboutit pas nécessairement à la réalisation d'un objectif chirurgical. Un ensemble d'étapes composent une phase chirurgicale, ou une tâche. Par exemple la tâche chirurgicale « Phacoémulsification » peut se décomposer en une succession d'étapes « Sillons » et « Cracking » qui permettent de « casser » le cristallin. Le terme de « **phases »** chirurgicales correspond dans la littérature, chez [21] ou [25] par exemple, à des tâches chirurgicales de haut niveau. Elles doivent aboutir à la réalisation d'un objectif chirurgical indispensable à la chirurgie. Une phase se termine quand l'objectif est atteint, cette phase n'aura donc plus lieu dans le reste de la chirurgie. Les phases sont donc présentes dans toutes les chirurgies, dans le même ordre. Enfin, le niveau de description le plus haut est la **procédure** chirurgicale elle-même. Ce niveau est utilisé dans le cas ou l'on cherche à différencier automatiquement le type de chirurgie, ou d'examen réalisé. C'est le cas de Twinanda et al., par exemple, qui cherchent à déterminer automatiquement le type de chirurgie abdominale pratiquée dans la vidéo requête [20] et de André et al. qui cherchent à fournir au praticien des exemples d'examens similaires [56].

Le choix du niveau de granularité dépend de l'objectif souhaité. Il n'est pas toujours nécessaire de décrire la chirurgie avec trop de précision. Néanmoins, dans notre cas, si nous souhaitons apporter une aide précise et être capable de reconnaître à l'avenir des situations délicates ou anormales, nous devons être capables d'analyser le processus chirurgical à un niveau suffisamment fin.

II.1.2 Les différentes descriptions des bases de données de vidéos médicales

Il existe dans la littérature différentes bases de données. En effet, puisqu'il n'existe pas de base de données commune, chaque équipe doit mettre au point sa propre base. La construction d'une base de données est une tâche complexe, particulièrement lorsqu'on souhaite travailler sur des données de chirurgies réelles et non des simulations. Cela requiert tout d'abord de collecter un grand nombre de données, puis de les annoter manuellement. Les annotations (typiquement la segmentation en phases ou en actions chirurgicales) doivent être pertinentes d'un point de vue chirurgical et permettre une reconnaissance automatique des cas similaires aisée. Les bases de données utilisées par les principales équipes du domaine de l'analyse automatique du processus chirurgical sont présentées ci-après et dans le Tableau 2.

II.1.2.1 Travaux de l'équipe VisAGeS

Très peu d'équipes ont travaillé sur la chirurgie de la cataracte, mais des travaux très intéressants ont été réalisés par F. Lalys de l'équipe MediCIS (Modélisation des connaissances et procédures chirurgicales et interventionnelles) de l'unité de recherche VisAGeS, sous la direction de P. Jannin [23]. Ces travaux ont porté sur l'analyse automatique de la chirurgie de la cataracte en s'appuyant sur une modélisation du processus chirurgical. Pour évaluer et valider leurs travaux, F. Lalys et P. Jannin ont travaillé sur une base de 20 chirurgies de la cataracte. Deux niveaux de description ont été utilisés. Dans un premier temps, les vidéos ont été décrites en **8 phases chirurgicales** : préparation, injection de Bétadine, incision de la cornée, hydrodissection, phacoémulsification, aspiration corticale, implantation de la lentille artificielle, ajustement de la lentille. L'enchaînement des phases chirurgicales est toujours le même et une phase commence lorsqu'une autre se termine. Puis, F. Lalys a essayé de travailler à un niveau de description plus fin : **18 activités** ont alors été identifiées. Une activité se définit par un triplet < verbe d'action – outil chirurgical – structure anatomique > : 12 verbes d'**action**, 13 **outils chirurgicaux** et 6 **zones d'action** ont ainsi été identifiés. A partir des 20 vidéos de la base de données de F. Lalys, **25 paires d'activités** ont été identifiées (une activité pour la main gauche, une activité pour la main droite). De très bons résultats ont été obtenus avec la description en phases chirurgicales, qui est assez proche de notre description en tâches chirurgicales, ce qui permet ensuite d'apporter une information contextuelle pour la reconnaissance automatique des activités chirurgicales. Cependant cette description comporte certaines limites, notamment parce qu'elle fusionne par exemple l'incision et l'injection du visqueux, ou l'hydrodissection et le capsulorhexis, qui sont des gestes différents d'un point de vue chirurgical. La description en paires d'activités, elle, permet de décrire la chirurgie avec plus de précision. Néanmoins, elle fournit des résultats plus inégaux en termes de reconnaissance automatique, notamment parce que certaines paires d'activités sont faiblement représentées dans la base de données.

II.1.2.2 Autres travaux

Il existe d'autres types de chirurgies sous contrôle vidéo, telles que les chirurgies laparoscopiques (ou cœlioscopies). Ces chirurgies sont réalisées à l'aide de petites incisons par lesquelles on fait passer une caméra endoscopique et les outils. Le chirurgien visualise donc la scène chirurgicale sur l'écran sur lequel sont retransmises les images. Ce type de chirurgie est essentiellement pratiqué sur l'appareil digestif, comme pour l'ablation de la vésicule biliaire par exemple. Plusieurs travaux ont été présentés autour des chirurgies laparoscopiques. Padoy et al. utilisent une base de données de 16 vidéos d'ablation de la vésicule biliaire [25]. La chirurgie se décrit en 14 phases chirurgicales. Tout comme la description en phases chirurgicales proposées par Lalys et al. [22], une phase commence lorsqu'une autre se termine et elles sont toujours réalisées dans le même ordre. Une phase contient des actions qui peuvent apparaître de façon répétitive dans chacune des phases. Ces actions sont fortement liées à l'utilisation des instruments. Ainsi un changement dans le type d'instruments utilisés représente un changement vers une nouvelle action.

Les robots dirigés par les chirurgiens tels que le robot Da Vinci, permettent également de réaliser des chirurgies sous contrôle vidéo, notamment des chirurgies laparoscopiques. Le robot Da Vinci comporte plusieurs bras manipulateurs dont un tient une caméra endoscopique. Ses bras sont manipulés à distance par un chirurgien qui visualise en direct la scène chirurgicale sur deux

écrans (un par œil). L'avantage de ce système dans le cadre de la reconnaissance automatique du geste chirurgical est qu'il fournit également, en plus de la vidéo, les trajectoires 3D des différents outils. Reiley et al. [37] et Zappella et al. [31] utilisent par exemple une base de 101 vidéos où chacune des vidéos représente une tâche chirurgicale. Ces tâches chirurgicales sont enregistrées dans le cadre de simulations et ne sont donc pas issues de chirurgies réellement pratiquées. Trois tâches différentes sont représentées : « suture » (39 essais), « passage de l'aiguille » (26 essais) et « nouage » (36 essais). Au sein de ces trois tâches chirurgicales, 15 actions ont été identifiées et annotées, telles que « passer l'aiguille », « positionner l'aiguille »... Les 15 actions ne sont pas nécessairement présentes dans chacune des tâches et une même tâche n'est pas toujours réalisée avec le même enchaînement d'actions.

Dans un cadre plus large, il existe d'autres sources de vidéos médicales, enregistrées dans le cadre d'examens ou des soins. Les examens endoscopiques sont par exemple une source importante de données vidéo. Il ne s'agit pas de procédures chirurgicales, néanmoins, l'analyse automatique de ces vidéos est similaire à l'analyse du processus chirurgicale. L'endoscopie est une méthode d'imagerie médicale permettant de visualiser des tissus ou des organes internes à l'aide une petite caméra qui retransmet les images à l'écran. Ces examens se font par les voies naturelles et les images sont interprétées en temps réel par les médecins, et peuvent être stockées dans le dossier des patients. De nombreux examens de ce type sont réalisés et le plus courant est la coloscopie, c'est-à-dire l'examen du côlon. L'utilisation de cette méthode d'imagerie est essentiellement pour le diagnostic, grâce à un examen visuel des tissus ou la réalisation de biopsies par exemple. Mais ils servent également à la thérapie, en réalisant l'ablation de tumeurs. Plusieurs travaux ont donc été développés autour de ce type d'examens. Stanek et al. travaillent par exemple avec 2464 h de vidéos de colonoscopie et de gastroscopie [18]. Cao et al. travaillent avec 25 vidéos de colonoscopies [19] et Mackiewicz et al. avec 76 vidéos de capsules endoscopiques sans fil [27]. Enfin André et al. [32] travaillent avec 118 vidéos de pCLE (Endomicrocopie Confocale par Minisondes).

Finalement, Droueche et al. [48] et Quellec et al. [49] ont, en plus de la base de chirurgie de la cataracte présentée dans le paragraphe II.2, travaillé avec une base constituée de chirurgies de pelage de la membrane épirétinienne.

Tableau 2. Bases de données utilisées pour les différents travaux en analyse automatiques de vidéos médicales

auteurs	Journal/conf	année	Taille de la base	Description
Leong et al.	MICCAI	2006	36 trajectoires (11 sujets)	4 Gestes laparoscopiques dans un simulateur
Cao et al.	IEEE TBE	2007	25 colonoscopies	2 Phases : Opération diagnostique/thérapeutique et transition
Mackiewicz et al.	IEEE TMI	2008	76 vidéos de capsules endoscopiques	4 phases : entrée, visualisation de l'estomac, de l'intestin, du colon
Reley et al.	MICCAI	2009	57 vidéos de sutures à 4 points avec le robot Da Vinci	9 sous-tâches chirurgicales

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

Lalys et al.	IEEE TBE	2012	20 chirurgies de la cataracte	12 phases chirurgicales
Andre et al.	IEEE TMI	2012	118 pCLE (endomicroscopie)	Procédures chirurgicales
Padoy et al.	Elsevier MIA	2012	16 chirurgies laparoscopiques	14 phases chirurgicales
Tao et al.	IPCAI	2012	101 séquences de 3 tâches chirurgicales séquences réalisée avec le robot Da Vinci	11 gestes chirurgicaux
Droueche et al.	IEEE EMBC	2012	69 séquences vidéos chirurgies de pelage la membrane épirétinienne; 1 séquence = 1 tâche chirurgicale	3 tâches chirurgicales
Staneck et al.	elsevier CMPB	2012	2464 h de vidéos à partir d'endoscopes (colonoscopie et gastroscopie)	2 phases : intérieur du patient, extérieur du patient
Lalys et al.	IEEE CARS	2013	20 chirurgies de la cataracte	8 phases chirurgicales et 25 paires d'activités
Zappella et al.	Elsevier MIA	2013	101 séquences de 3 tâches chirurgicales séquences réalisée avec le robot Da Vinci	15 gestes chirurgicaux
Oropesa et al.	Surg Endosc	2013	42 gestes laparoscopiques dans un simulateur	
Tao et al.	MICCAI	2013	101 séquences de 3 tâches chirurgicales séquences réalisée avec le robot Da Vinci	15 gestes chirurgicaux
Garraud et al.	Surgetica	2014		Ontologie du processus chirurgicale
Quellec et al.	MIA	2014	69 séquences vidéos de chirurgies de pelage de la membrane épirétinienne; 900 séquences vidéos de chirurgies de la cataracte ; 1 séquence = 1 tâche chirurgicale	3 tâches chirurgicales de chirurgies de pelage de rétine et 9 tâches chirurgicales de chirurgies de la cataracte

Quellec et al.	TMI	2014	186 vidéos de chirurgies de la cataracte	9 tâches chirurgicales
Suzuki et al.	Surg Endosc	2014	Gestes laparoscopiques dans un simulateur : 6 experts et 11 novices. 5 sutures par chirurgien	2 niveaux d'expérience : expert et novice
Forestier et al.	CARS	2015	22 chirurgies d'hernies discales lombaires	4 phases chirurgicales et 23 000 paires d'activités
Quellec et al.	TMI	2015	900 séquences vidéos de chirurgies de la cataracte ; 1 séquence = 1 tâche chirurgicale	9 tâches chirurgicales
Twinanda et al.	CARS	2015	208 vidéos représentant 8 chirurgies laparoscopiques différentes	8 chirurgies laparoscopiques

Ces différents exemples de bases de données et de descriptions du processus chirurgical ont permis de nourrir une réflexion sur une manière pertinente et complète de décrire une chirurgie. Lalys et al. [21] et Forestier et al. [26] ont montré l'intérêt de travailler à différents niveaux de granularité. Lalys et al. utilisent l'information issue de la reconnaissance plus aisée des tâches chirurgicales pour apporter une information contextuelle lors de la reconnaissance automatique des activités chirurgicales [21]. A l'inverse, Forestier et al., utilisent comme observation des activités chirurgicales réalisées, pour reconnaître automatiquement les phases chirurgicales [26]. Cette approche multi-échelles semble particulièrement intéressante. Il semble également intéressant d'utiliser les avantages des différents niveaux pour les faire travailler ensemble pour l'analyse automatique du processus chirurgical.

II.2 La base de données du LaTIM

Comme il n'existe pas de bases de données publiques de vidéos médicales pour évaluer nos méthodes, le LaTIM a constitué une base de données de vidéos de chirurgie de la cataracte, grâce à une collaboration avec le service d'ophtalmologie du CHRU de Brest.

II.2.1 Application clinique : la chirurgie de la cataracte

La cataracte est une maladie due à une opacification progressive du cristallin, c'est-à-dire de la lentille qui forme les images sur la rétine (Figure 13). Le cristallin est normalement transparent et son opacification entraîne une baisse de la vision. La cataracte est majoritairement due à l'âge et, en France, elle touche 20 % de la population après 65 ans et plus de 60 % des personnes après 85 ans[3].

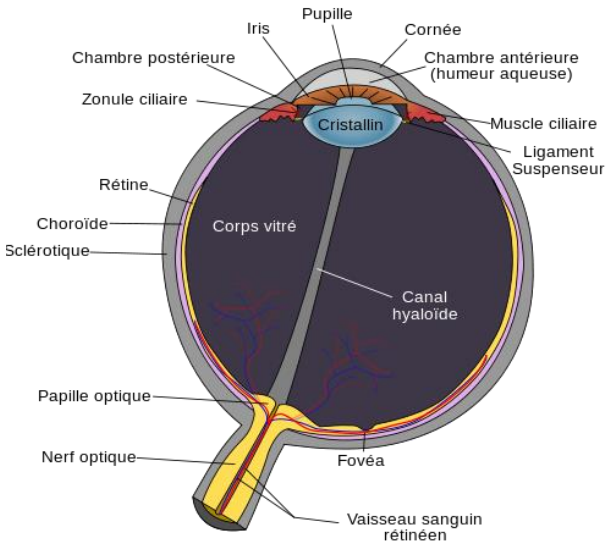


Figure 13. Structure anatomique de l'œil humain

Lorsque l'opacification devient trop gênante pour le patient, le seul traitement est la **chirurgie**. La procédure consiste à retirer le cristallin opacifié. Deux petites incisions sont réalisées pour atteindre le cristallin (une incision principale et une incision secondaire également appelée contre incision). Un gel visqueux est injecté pour conserver la pression dans l'œil. Puis une ouverture dans la poche qui le contient (sac cristallin) est réalisée pour retirer ensuite le cristallin. Cette poche est conservée pour recevoir ensuite un cristallin artificiel (implant intraoculaire). Une fois l'implant en place, le visqueux est retiré puis les incisions sont refermées par hydro-suture ou par un fil.

L'intervention est généralement pratiquée sous anesthésie locale et en ambulatoire : la personne arrive le matin, est opérée et peut sortir le jour même, si elle est accompagnée. L'opération dure en moyenne 15 minutes et avec environ 700 000 interventions par an, elle est actuellement

[3] http://www.ameli-sante.fr

la chirurgie la plus pratiquée en France3. L'intervention est pratiquée sous **contrôle oculaire, avec renvoi vers un système d'acquisition vidéo**. Un microscope binoculaire chirurgical est utilisé par le chirurgien pour pratiquer l'intervention. Ce microscope génère deux flux optiques analogiques. Chacun est découpé en deux (en parts non égales). Les deux flux principaux vont vers le chirurgien, un troisième va vers l'aide opératoire (moniteur vidéo) et le quatrième vers la caméra pour être enregistré. La chirurgie de la cataracte est très pratiquée, très codifiée et reproductible ; elle permet donc de développer et d'évaluer des outils d'aide à la chirurgie, s'appuyant sur une recherche par le contenu de cas similaires, grâce aux enregistrements vidéos, sans nécessiter une base de données trop importante. En effet une chirurgie plus diversifiée nécessiterait plus de données pour couvrir tous les cas possibles. C'est pourquoi nous avons choisi de travailler sur cette chirurgie pour valider nos méthodes. Néanmoins, il existe d'autres bases de vidéos de chirurgies dans la littérature.

II.2.2 Les données

La constitution d'une base de données conséquente est donc une étape essentielle, mais complexe, pour le développement et l'évaluation de nos méthodes de recherche de vidéos par le contenu. Entre février et juillet 2011, 186 vidéos ont été enregistrées et collectées au service d'ophtalmologie du Centre Hospitalier Régional et Universitaire (CHRU) de Brest. Ces chirurgies ont été réalisées par 10 chirurgiens différents dans deux blocs opératoires différents avec des systèmes d'acquisition des vidéos différents. Dans la première salle, les vidéos ont été enregistrées au format DV à l'aide d'un système CCD-IRIS (Sony, Tokyo, Japon) associé à un système d'enregistrement vidéocassette DSR-20MDP (Sony, Tokyo, Japon). Dans la seconde salle, les vidéos ont été enregistrées au format MPEG 2 à l'aide d'un système d'enregistrement MediCap USB200 (MediCapture, Philadelphie, USA). Aucune sélection particulière n'a été effectuée. Il s'agit donc d'une vraie base de données terrain. La base de données est complète et variée, que ce soit au niveau des patients opérés, du type de cataracte (plus ou moins avancée) ou des chirurgiens (internes débutants ou intermédiaires, séniors, etc…). Certaines vidéos peuvent être incomplètes, ou contenir des cas particuliers, comme la pose d'écarteurs d'iris ou des petites complications comme l'échec au premier essai de la pose de l'implant, ou l'apparition d'hernies de l'iris. Les durées d'exécution des chirurgies et des différents gestes chirurgicaux sont également très variées. Par exemple, le geste chirurgical qui consiste à réaliser des sillons pour faciliter la mise en morceaux du cristallin a une durée moyenne de 43 secondes avec un écart type de 42 secondes au sein de notre base de données. De même, la tâche chirurgicale incision a une durée moyenne de 1'17 minutes, avec un écart type de 0'17 minutes.

II.2.2.1 Description de la chirurgie

Les 186 vidéos ont été annotées dans un premier temps en se basant sur une description en 10 tâches chirurgicales, définies par les chirurgiens. La description en 10 tâches chirurgicales utilisée pour décrire l'intégralité de la base de données se compose des 9 tâches principales présentées dans la Figure 14 (Incision, Rhexis, Hydrodissection, Phacoémulsification, Epinoyau, Injection visqueux, Mise en place, Retrait du visqueux, Fermeture) plus une tâche « Divers ». La tâche « Incision » consiste en la réalisation des deux petites incisions. Le Rhexis est la réalisation d'une ou-

verture circulaire dans la poche du cristallin. Une fois cette ouverture réalisée, le cristallin est décollé de son enveloppe (Hydrodissection) avant d'être séparée en morceaux et aspiré (Phacoémulsification et Epinoyau). La poche peut alors recevoir le cristallin artificiel (Mise en place). Le visqueux est ensuite retiré (Retrait du visqueux) et remplacé par un liquide physiologique avant la Fermeture. L'injection du visqueux permet de conserver la pression dans l'œil, cette tâche est réalisée plusieurs fois au cours de la chirurgie, notamment au moment de l'incision, et avant l'injection de l'implant. La tâche « Divers » regroupe toutes les tâches non classiques, c'est-à-dire particulières à certaines chirurgies de la cataracte. Par exemple, dans le cas où la pupille n'est pas suffisamment dilatée, des écarteurs peuvent être installés, ou pour certains implants corrigeant l'astigmatie, des mesures d'angles sont effectuées. Pour chaque vidéo, les chronométrages de début et de fin de chacune des tâches ont été annotés. Il a été considéré pour cette annotation qu'une tâche commence lorsque le premier instrument correspondant à cette tâche entre dans le champ de vue de la caméra et qu'elle se termine lorsque le dernier instrument sort du champ de vue.

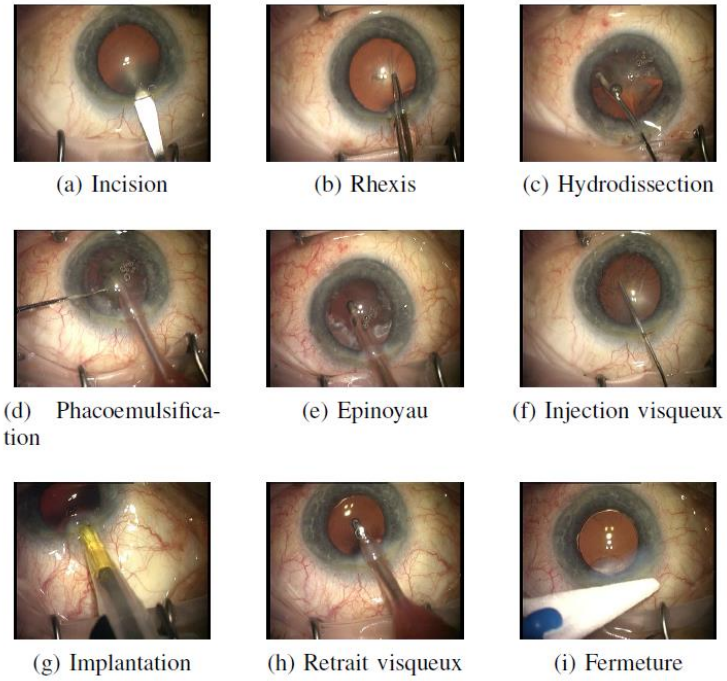


Figure 14. Les neufs tâches chirurgicales principales

Cette description a permis de tester et de valider un grand nombre des algorithmes développés. Cependant elle ne permet pas de décrire la chirurgie de façon suffisamment précise et complète. Il peut par exemple être intéressant de différencier la pose d'écarteurs d'iris de la mesure d'angle, alors que ces deux étapes sont regroupées actuellement dans une unique tâche « divers ». De même la tâche « incision » regroupe la réalisation de l'incision principale et de l'incision secondaire. Or le visqueux est généralement injecté entre ces deux incisions. Comme une tâche commence lorsque le premier instrument correspondant à cette tâche entre dans le champ de vue de la camera et qu'elle se termine lorsque le dernier instrument en sort, la tâche « Injection visqueux » chevauche la tâche « Incision ». Cela peut rendre compliqué le séquençage automa-

tique de la chirurgie. De plus, certaines recommandations ou alertes peuvent nécessiter de connaître avec plus de précision ce que le chirurgien est en train d'effectuer, par exemple, quel geste chirurgical est en train d'être effectué. En revanche, dans certains cas, un niveau de description plus grossier est suffisant. Nous avons donc travaillé avec David Martiano, un interne en chirurgie au CHRU de Brest, à une nouvelle description multi-échelles [57] (paragraphe II.3).

II.2.2.2 Les bases de données annotées

De nombreuses informations nous ont été apportées par les chirurgiens pour chacune des vidéos. Nous connaissons par exemple l'âge des patients, le type d'implant utilisé, l'expérience du chirurgien qui opère, etc… Mais surtout, les chirurgiens, en s'appuyant sur la description présentée précédemment (paragraphe II.2.2.1) ont annoté manuellement les vidéos. Cette étape d'annotation manuelle des vidéos a permis d'aboutir à la création de plusieurs bases de données vidéo. Une base initiale de **186 chirurgies décrites en tâches chirurgicales** et a ainsi été constituée dans un premier temps.

Une **sous-base de 30 vidéos** sélectionnées dans la base de données initiale a également été constituée. Cette sous-base a été construite par David Martiano, un interne en chirurgie au CHRU de Brest, et moi-même, pour être représentative de la base de données principale afin d'être réinterprétée selon une description multi-échelles de la chirurgie (paragraphe II.3). Des chirurgiens de niveaux d'expérience différents (débutants, intermédiaires et sénior) sont représentés et trois cas de mesure d'angle sont également présents dans la sous-base. Cette base nous a servi de support pour évaluer l'ensemble des méthodes présentées dans cette thèse.

II.3 Nouvelle description multi-échelles de la chirurgie

Si nous souhaitons apporter des informations précises et pertinentes aux chirurgiens pendant sa chirurgie, il est nécessaire de reconnaître à chaque instant quel geste chirurgical est en train d'être réalisé. Cela implique une analyse fine de la chirurgie en cours d'exécution. Cependant, travailler à un niveau de granularité trop fin augmente les difficultés en matière de reconnaissance automatique [23]. A l'inverse, utiliser un niveau de granularité élevé permet une reconnaissance plus aisée, mais ne permet pas de décrire la chirurgie avec assez de précision. Nous avons donc réfléchi, avec David Martiano, un interne en chirurgie du service d'ophtalmologie du CHRU de Brest, a une nouvelle description de la chirurgie de la cataracte.

Nous avons décidé de décrire la chirurgie avec 3 nouveaux niveaux de description. Ces trois niveaux de descriptions permettent de décrire la chirurgie en allant d'un niveau de description fin et précis jusqu'à un niveau beaucoup plus général. Les niveaux choisis sont les activités (composées d'un instrument et d'un geste), les étapes et les phases. Grâce aux quatre niveaux de description ainsi mis en place, nous aurons alors la description la plus complète possible de la chirurgie, en allant d'un niveau de description fin et précis jusqu'à un niveau beaucoup plus général. En effet, une activité est composée d'un instrument et d'un geste, et il n'est pas nécessaire d'utiliser le niveau « procédure » puisque l'on n'a qu'un type de procédure dans notre base de données de chirurgie. Compte tenu des temps d'annotation plus longs, une sous-base de 30 vidéos a été annotée manuellement avec cette nouvelle description.

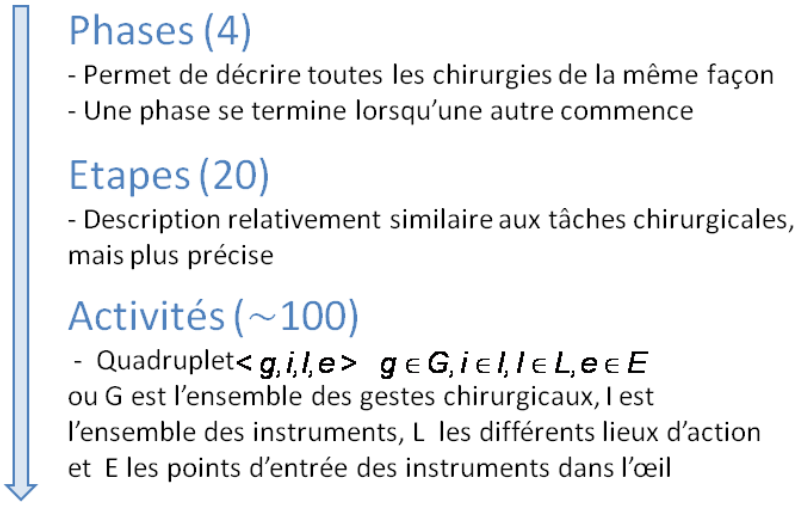


Figure 15. Les trois nouveaux niveaux de granularités

II.3.1 Phases

La description en phases de la chirurgie de la cataracte est la description la plus générale. Nous utilisons la définition d'une phase chirurgicale proposée par Padoy et al. [25]: Une phase chirurgicale est une partie de la chirurgie qui peut être identifiée de façon unique dans tous les

exemples de chirurgie et qui est validée par un chirurgien expert. De plus, nous considérons qu'une phase doit aboutir à la réalisation d'un objectif chirurgical indispensable à la chirurgie. Ainsi quatre phases chirurgicales ont été identifiées : 1 - l'ouverture, 2 - la phacoémulsification, 3 - l'implantation et 4 - la fermeture, plus une phase « Transition » qui comprend les débuts et fins des vidéos, dans lesquels il ne se passe rien de pertinent. Les phases s'enchaînent toujours de la même manière et une phase se termine quand une autre commence. Cette description en phases permet de décrire toutes les chirurgies de la cataracte de la même manière, avec le même enchaînement de phases quel que soit le type de chirurgie de la cataracte ou le chirurgien qui la pratique.

(a) Rien (b) Ouverture (c) Phacoemulsification (d) Implantation (e) Fermeture

Figure 16. Les phases de la chirurgie de la cataracte

II.3.2 Etapes

La description en étapes chirurgicales est un niveau de description intermédiaire. Vingt étapes chirurgicales ont été identifiées. Cette description se rapproche de celle en tâches chirurgicales, mais elle permet par exemple de différencier les incisions principales et secondaires. De plus, elle prend en compte les étapes spécifiques à certaines opérations de la cataracte telle que la pose d'écarteurs d'iris, ou la mesure d'angles pour les implants corrigeant l'astigmatie. Il s'agit donc d'une description plus détaillée que la description en tâches, sans pour autant être aussi précise et donc complexe que la description en activités. Comme cela est présenté dans la Figure 17, les phases chirurgicales sont constituées d'un ensemble d'étapes chirurgicales.

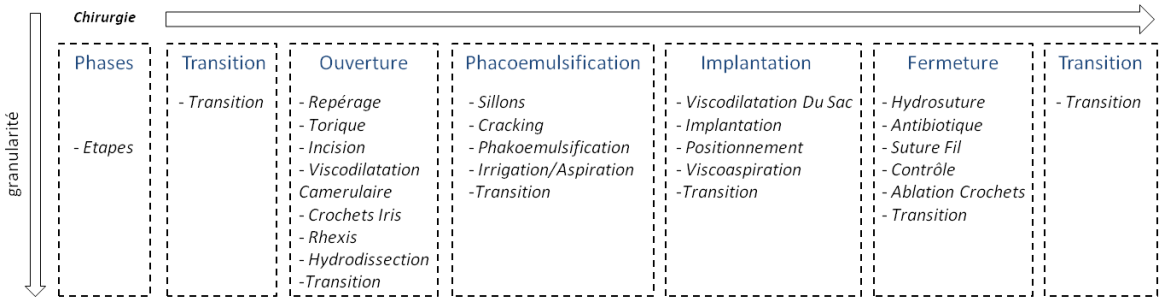


Figure 17. Relations entre les différentes étapes et phases qu'il est possible de rencontrer dans une chirurgie de la cataracte selon la description de David Martiano et moi-même

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

II.3.3 Activités

La description en activités est la plus précise. Elle permet de décrire chaque geste chirurgical, effectué par chacune des mains du chirurgien ou de l'aide, en un lieu donné. Elle représente tout ce qui se passe dans une chirurgie. Nous nous sommes inspirés du travail de Lalys et al. [21] pour définir nos activités, en ajoutant une quatrième information relative au point d'entrée. En effet, deux activités peuvent toutes deux avoir lieu dans la même zone anatomique de l'œil, la chambre antérieure (CA) par exemple, mais l'outil peut être inséré soit par l'incision principale, soit par la contre-incision. L'utilisation de cette information dans la description permet une construction plus précise de notre lot d'activités, en localisant avec plus de précision la zone d'action des outils. Les activités sont ainsi décrites par un quadruplet :

$$< g, i, l, e > \; g \in G, i \in I, l \in L, e \in E$$

ou G est l'ensemble des gestes chirurgicaux présents dans notre base de données, I est l'ensemble des instruments utilisés pour la chirurgie de la cataracte, L les différents lieux d'action et E les points d'entrée. 68 gestes chirurgicaux ont été identifiés par David Martiano, ainsi que 34 instruments, 10 lieux d'action et 3 points d'entrée. Cela a permis d'identifier 87 activités différentes au sein de notre base de données.

II.4 Diagrammes de transition obtenus

Afin de visualiser les différents déroulements possibles de la chirurgie, nous avons construit des diagrammes de transition pour chacun des trois nouveaux niveaux ainsi que pour la description en tâches chirurgicales, selon le processus présenté dans la Figure 18. Ces diagrammes nous serviront également pour la mise au point d'outils de suivi automatique de la chirurgie.

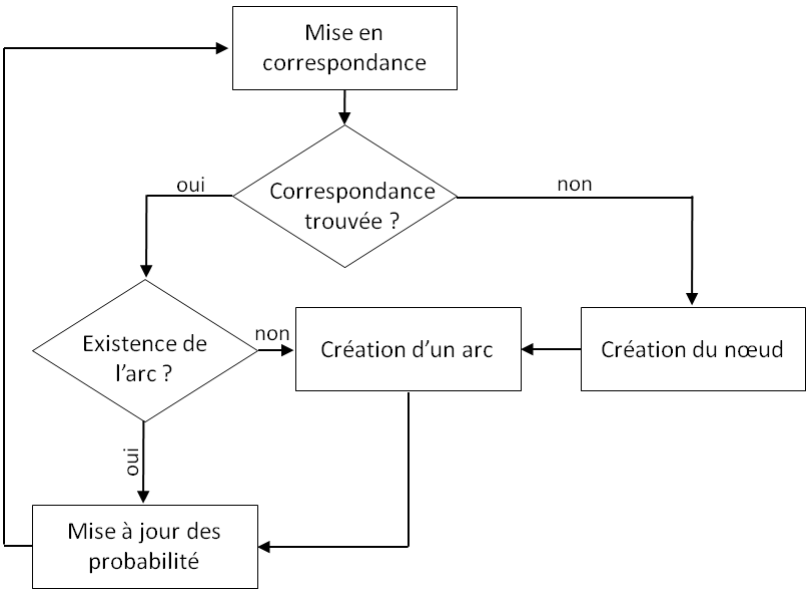


Figure 18. Processus de construction de graphes pour la modélisation du processus chirurgical

Ces diagrammes de transition permettent de représenter visuellement le déroulement du processus chirurgical aux différents niveaux de description. Chaque nœud représente une tâche (ou phase, étape, etc…). Lorsque deux tâches chirurgicales ont lieu en parallèle (comme l'incision et l'injection du visqueux, par exemple), une nouvelle tâche est créée, qui est une combinaison des deux autres tâches. Les différents arcs (S_i, S_j) représentent la probabilité d'avoir la tâche S_j lorsque S_i prend fin. Lorsque cette probabilité est inférieure ou égale à $0{,}1$, l'arc est représenté en pointillé.

II.4.1 Phases

Le diagramme de transition des phases chirurgicales, obtenu à partir des 30 vidéos de la base nouvellement interprétée, est présenté dans la Figure 19. Nous pouvons constater que les quatre phases principales plus la phase « Transition » s'enchaînent toujours de la même manière. De plus une phase termine quand une autre commence. Ainsi le diagramme est très simple, ce qui devrait rendre plus facile l'analyse du processus chirurgical à ce niveau.

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

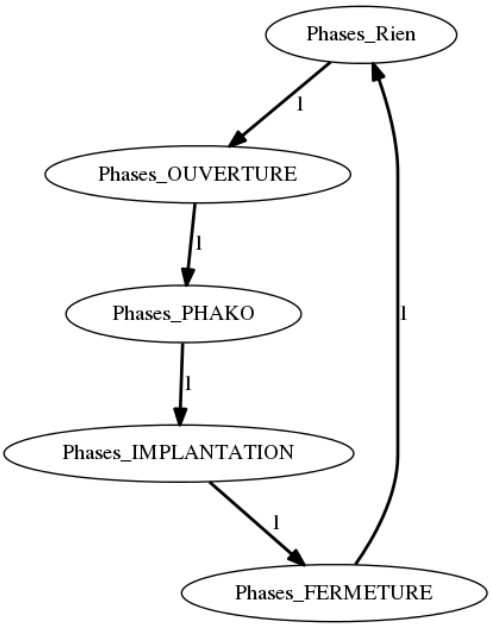


Figure 19. Diagramme de transition des phases chirurgicales

II.4.2 Tâches

Le digramme obtenu, avec les mêmes vidéos, pour les tâches chirurgicales est présenté dans la Figure 20.

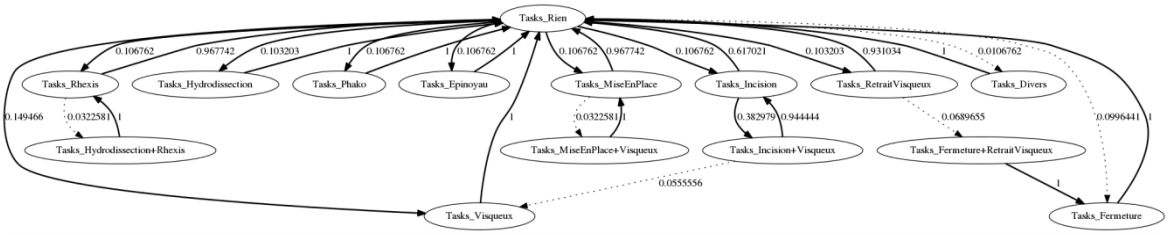


Figure 20. Diagramme de transition des tâches chirurgicales

Une première constatation est que la tâche « Rien » s'intercale entre quasiment chacune des tâches chirurgicales. Ceci montre l'intérêt de la méthode proposée par Quellec et al. qui consiste à détecter ces transitions [51]. Nous constatons par ailleurs, que le diagramme, bien que plus complexe que le diagramme des phases, reste simple et facilement interprétable visuellement.

II.4.3 Etapes

Le diagramme obtenu pour la description en étapes chirurgicales est présenté en Annexe 2. Un zoom sur l'organisation des étapes appartenant à la phase phacoémulsification est présenté dans la Figure 21.

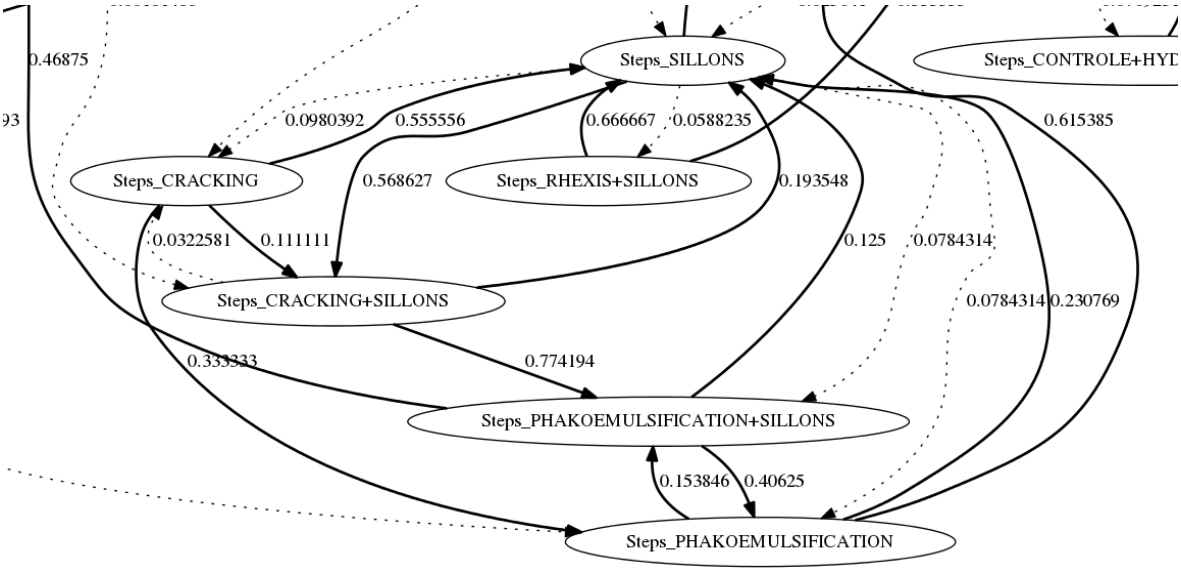


Figure 21. Zoom sur l'organisation des étapes appartenant à la tâche phacoémulsification du diagramme de transition des étapes chirurgicales

On constate la complexité du processus chirurgical à ce niveau de description par rapport à la description en tâches chirurgicales. Notamment, on constate qu'il existe de nombreux allers et retours entre les différentes étapes qui permettent d'aboutir à la réalisation d'une tâche chirurgicale telle que la phacoémulsification. En effet, plusieurs étapes « Sillons » et « Cracking » sont réalisées consécutivement afin de réduire le cristallin en morceaux. Dans ce cas, les étapes chirurgicales ne sont pas séparées par une étape de transition. La méthode proposée par Quellec et al. [51] semble alors difficilement adaptable à ce niveau de description.

II.4.4 Activités

Tout comme pour la description étapes chirurgicales, un focus sur l'organisation des activités appartenant à la phase phacoémulsification est présenté dans la Figure 22. On constate que l'organisation des activités est extrêmement complexe. Il existe un grand nombre d'activités et d'ordonnancements possibles. Le diagramme complet est complexe est difficilement interprétable visuellement.

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

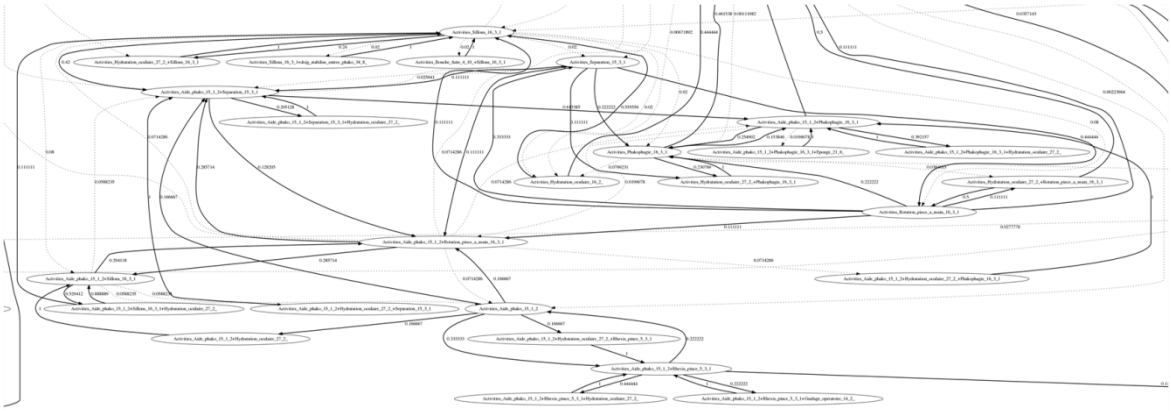


Figure 22. Zoom sur l'organisation des activités appartenant à la tâche phacoémulsification du diagramme de transition des activités chirurgicales

II.5 Discussion - Conclusion

Grâce à une forte collaboration avec le service d'ophtalmologie du CHRU de Brest, le LaTIM dispose de bases de données importantes, annotées par des médecins. Une base de 186 chirurgies de la cataracte, décrite en tâches chirurgicales, a ainsi été constituée. Une sous-base de 30 vidéos de chirurgies a également été construite à partir d'une description multi-échelles. Ces bases représentent une grande variété de cas, et de chirurgiens, et sont ainsi représentatives des chirurgies de la cataracte qui ont lieu quotidiennement dans les blocs opératoires du CHRU de Brest.

Les trois nouveaux niveaux de description définis permettent de décrire de façon précise et complète la chirurgie. Ils permettent de travailler à différents niveaux de précision pour l'analyse et la modélisation de la chirurgie de la cataracte et d'exploiter ainsi les avantages de chacune des descriptions. Les alertes et les recommandations pourront ainsi être ciblées. Les diagrammes de transition obtenus pour chacune des étapes permettent de se rendre compte de la grande variabilité de réalisation des différents niveaux de description, particulièrement pour les étapes et les phases. Pour ces deux niveaux, la reconnaissance automatique sera donc beaucoup plus complexe que pour les niveaux de granularité plus élevés, tels que les tâches ou les phases chirurgicales. C'est pourquoi nous souhaitons tirer avantage des différents niveaux en les faisant travailler ensemble pour permettre le séquençage automatique d'une vidéo de chirurgie. Les diagrammes de transitions ainsi construits introduisent l'idée de s'appuyer sur des modèles statistiques du processus chirurgical pour aider au séquençage automatique des vidéos.

Pour conclure, nous avons trois nouveaux de description qui permettent de décrire la chirurgie avec beaucoup de précision et 30 vidéos de chirurgies de la cataracte interprétées. Il est maintenant nécessaire de s'atteler aux problèmes de reconnaissance automatique du geste chirurgical.

Dans le paragraphe suivant, nous chercherons dans un premier temps à reconnaître automatiquement la tâche réalisée dans la séquence présentée en requête. Nous étudierons les méthodes de recherche de cas similaires pour la reconnaissance automatique des tâches chirurgicales. Les méthodes de caractérisation et de catégorisation choisies seront ensuite évaluées sur une base de séquences vidéo, chaque séquence représentant une tâche chirurgicale.

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

Chapitre III. Reconnaissance automatique de tâches chirurgicales

III.1 Indexation et Recherche de vidéos par le contenu (CBVR) 67
III.1.1 Caractérisation des vidéos 67
III.1.1.1 Analyse de la structure 68
III.1.1.1.1 Images 68
III.1.1.1.2 Sous-séquences 69
III.1.1.2 Extraction de caractéristiques visuelles 69
III.1.1.2.1 Caractéristiques statiques : couleurs, textures, formes 69
III.1.1.2.2 Caractéristiques dynamiques : Objets, Mouvement 70
III.1.1.3 Construction des signatures visuelles 71
III.1.2 Mesure de similitude 73
III.1.2.1 Alignement dynamique temporel (DTW) 73
III.1.2.2 Mesure de similitude de Piciarelli et al. 74
III.1.3 Catégorisation des vidéos 75
III.1.4 Synthèse 76

III.2 Caractérisation des vidéos : nos choix 77
III.2.1.1 Extraction de caractéristiques basées sur le mouvement 77
III.2.1.1.1 Histogrammes de mots visuels 77
III.2.1.1.2 Histogrammes de Mouvement 78
III.2.1.2 Normalisation des vidéos 79

III.3 Catégorisation des vidéos : nos choix 82
III.3.1 Mesure de similitude de Piciarelli et al. 82
III.3.2 Recherche des plus proches voisins 83

III.4 Evaluation 85
III.4.1 La base de sous-séquences vidéos 85
III.4.2 Mesure de la performance : l'aire sous la courbe ROC 85
III.4.3 Résultats 87
III.4.3.1 Mesure de similitude de Piciarelli et al. 87
III.4.3.2 Influence du choix de la caractérisation des vidéos 89
III.4.3.3 Temps de calcul 93

III.5 Discussion – Conclusion 94

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

Nous avons vu au chapitre I que l'utilisation des méthodes de recherche de vidéos par le contenu (CBVR) offrait une première réponse à l'analyse automatique de vidéos. Ces méthodes permettent de chercher les cas les plus proches des vidéos (ou séquences vidéo) présentées en requête. En fonction des tâches représentées parmi les cas les plus proches retrouvés, on peut estimer la tâche la plus probablement présente dans la séquence requête. Nous présentons plus en détail dans ce chapitre les méthodes de recherche de cas similaires existantes dans la littérature et plus particulièrement les méthodes de caractérisation et de comparaison des vidéos. Les méthodes de caractérisation et de catégorisation choisies seront évaluées sur une base de séquences vidéo, où chaque séquence représente une tâche chirurgicale. Nous chercherons alors à reconnaître automatiquement la tâche réalisée dans la séquence présentée en requête.

III.1 Indexation et Recherche de vidéos par le contenu (CBVR)

Les méthodes de recherche de vidéos par le contenu s'appuient sur la comparaison d'une vidéo requête avec les cas de la base de données en comparant leur contenu visuel. Le principe général de la CBVR est présenté dans le paragraphe I.1.2.3 (page 27) et comporte deux aspects clés : la caractérisation des vidéos par leur contenu visuel et la comparaison du cas requête avec les cas de la base de données. De par la complexité du contenu qu'elle représente, la CBVR est plus complexe que la CBIR. En effet, l'aspect temporel s'ajoute à l'aspect spatial, avec une complexité supplémentaire due au fait que les vidéos sont généralement de durées différentes, avec des actions dont les vitesses d'exécution sont également variables. Cet aspect est donc important à prendre en compte. Il existe, dans la littérature, différentes méthodes pour caractériser les vidéos et les comparer, en fonction des applications souhaitées. Elles s'appliquent cependant le plus souvent à des vidéos complètes (analyse après acquisition).

III.1.1 Caractérisation des vidéos

La première étape de la CBVR consiste à représenter les vidéos par des descripteurs qui sont interprétables par un ordinateur. Cette étape est également appelée indexation. Ces descripteurs sont des vecteurs de caractéristiques, appelés signatures visuelles. Cette étape est importante, car il est nécessaire de bien représenter le contenu des vidéos, souvent complexe, pour pouvoir ensuite les comparer aisément aux cas de la base de données.

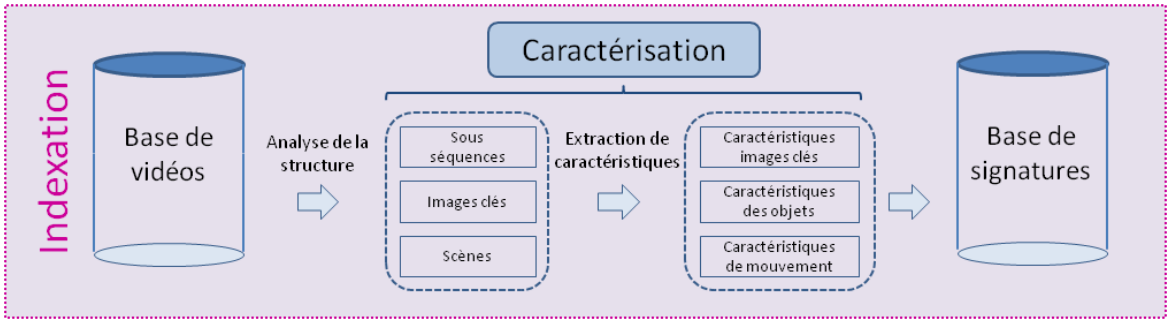


Figure 23. Caractérisation des vidéos par leur contenu visuel

Le principe général de la caractérisation des vidéos est présenté dans la Figure 23. Une première étape de la caractérisation consiste à analyser la structure de la vidéo pour en extraire des scènes, sous-séquences ou images clés à partir desquelles seront extraites les caractéristiques visuelles.

III.1.1.1 Analyse de la structure

Les vidéos sont une succession d'images ou de trames qui peuvent être entrelacées. Il y a donc différentes manières d'extraire l'information pour caractériser une vidéo. On peut notamment extraire des caractéristiques visuelles dans chacune des images, ou uniquement au sein d'images clés, pour ensuite construire une signature visuelle par image, par sous-séquence, par scène clé ou par vidéo. Il est donc nécessaire de bien choisir le type d'analyse de structure à utiliser, en fonction de l'application voulue. La Figure 24 présente les différentes façons d'utiliser la structure de la vidéo pour extraire les signatures visuelles.

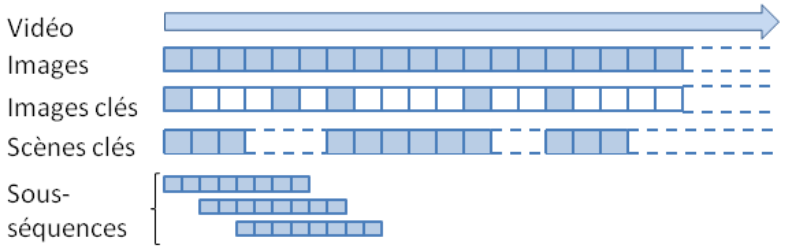


Figure 24. Différentes façon d'utiliser la structure de la vidéo pour l'extraction des caractéristiques visuelles

III.1.1.1.1 Images

Une première idée simple consiste à se ramener à un problème de classification d'images, en cherchant les plus proches voisins de chacune des images ou images clés pour les catégoriser. F. Lalys a évalué cette approche dans sa thèse [23] en cherchant à classifier indépendamment chacune des images pour la reconnaissance automatique des phases chirurgicales. Les taux de bonne classification obtenus sont de l'ordre de 82 %. Cependant des confusions ont été observées entre des étapes qui n'ont pas lieu au même moment de la chirurgie (comme l'incision et la suture par exemple). Cela est dû au fait que l'information temporelle n'est pas prise en compte. Fusco et al.

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

[58] proposent dans le cadre de la reconnaissance automatique de visages, un algorithme pour lequel une recherche des plus proches voisins est effectuée pour chaque image d'une sous-séquence $[t-T, t]$ de la vidéo. Les résultats obtenus pour chacune des images de la sous-séquence sont combinés pour calculer le score de similitude de l'image enregistrée au temps t.

Une autre approche consiste à extraire des signatures visuelles dans chacune des images, ou images clés qui composent la vidéo, pour construire la signature visuelle de la vidéo complète. Cette approche a par exemple été utilisée par André et al., qui construisent une signature par vidéo à partir des sacs de mots visuels extraits dans chacune des images [32].

III.1.1.1.2 Sous-séquences

Les vidéos peuvent également être caractérisées en s'appuyant sur un découpage en sous-séquences de la vidéo. Quellec et al., proposent par exemple, dans le système décrit dans le paragraphe I.2.2.1, d'utiliser de courtes séquences extraites du flux vidéo. [49]. Ces séquences sont ensuite caractérisées puis comparées à des sous-séquences archivées. Su et al. divisent également leurs vidéos en un ensemble de sous-séquences [59]. Chaque sous-séquence est ensuite représentée par une image clé à partir de laquelle sont extraites les caractéristiques visuelles. Certaines séquences sont alors éliminées pendant que d'autres sont regroupées et encodées pour construire un ensemble de motifs temporels qui représenteront la vidéo.

L'analyse de la structure est donc une étape à prendre en compte avant d'extraire les caractéristiques visuelles pour représenter le contenu des vidéos, souvent complexe, de façon pertinente.

III.1.1.2 Extraction de caractéristiques visuelles

Différents types de caractéristiques visuelles peuvent être extraits à partir des images ou séquences d'images qui composent la vidéo, tels que les formes, les couleurs ou le mouvement par exemple.

III.1.1.2.1 Caractéristiques statiques : couleurs, textures, formes

L'extraction des **couleurs** pour caractériser les vidéos s'inspire directement de l'analyse d'images. L'extraction des couleurs dépend du type de représentation utilisé. Le plus courant est la représentation RGB (Rouge, Vert, Bleu ou « Red, Green, Blue » en anglais). La couleur d'un pixel est alors une combinaison linéaire des trois couleurs primaires : le *rouge*, le *vert* et le *bleu*. Mais il existe de nombreuses autres représentations, telles que la représentation HSV (Teinte, Saturation, Valeur ou « Hue Saturation Value » en anglais) ou YUV (Y représente la luminance et U et V la chrominance). La *teinte* est la mesure de l'angle sur un cercle des couleurs, la *saturation* est la pureté de la couleur et la *valeur* est la brillance. La *luminance* représente l'information de luminosité (similaire à la brillance) alors que la *chrominance* représente l'information de couleur. Le choix de l'espace des couleurs dépend de l'application souhaitée. Blum et al. utilisent par exemple les 3 canaux RGB et les 3 canaux HSV pour construire leurs signatures visuelles [24]. De même, Twinanda et al. utilisent les canaux *rouge-vert-bleu* de la représentation RGB et *teinte-saturation* de la représentation HSV [20]. Dans le domaine médical, Fisher et Mackiewicz proposent une revue

bibliographique des différentes méthodes d'analyse de couleurs dans le cadre des vidéos d'examens par capsules endoscopiques [28]. Cependant, dans le cas de la chirurgie de la cataracte, les couleurs varient peu au cours du temps et ne sont pas suffisantes pour différencier les différentes phases chirurgicales. Lalys et al. utilisent néanmoins, entre autres caractéristiques visuelles, les variations de couleurs de la pupille [22], et Quellec et al. utilisent des informations de couleurs, associées à des informations de texture et de mouvement [60].

Tout comme l'utilisation des couleurs, l'utilisation des **textures** pour caractériser les vidéos s'inspire directement de l'analyse d'images. Les textures représentent des caractéristiques homogènes, telles que des motifs répétitifs ou des caractéristiques fréquentielles. Différentes approches permettent de représenter l'information de texture dans une image. Ramamurthy et al. [61], par exemple, utilisent les matrices de cooccurrences pour leur algorithme de CBIR. Mackiewicz et al. utilisent quant à eux les motifs binaires locaux (ou LBP pour « Local Binary Pattern » en anglais) pour extraire des caractéristiques de textures dans les vidéos de capsules endoscopiques [27]. Enfin, Quellec et al. utilisent la transformée en ondelettes [60]. Ces caractéristiques sont généralement associées à des informations de couleur pour construire les signatures visuelles.

Un autre type d'informations peut être extrait des images pour construire les signatures visuelles : les informations de **formes**. Ces informations sont souvent obtenues par un détecteur de contour tel que l'application d'un gradient horizontal et vertical. Ces informations sont utilisées dans les descripteurs SIFT (transformation de caractéristiques visuelles invariantes à l'échelle, « Scale-Invariant Feature Transform » en anglais) utilisés par [56] par exemple.

Ces caractéristiques sont statiques et permettent de caractériser les images qui composent la vidéo. Cependant elles ne prennent pas en compte l'aspect dynamique qu'apporte la vidéo par rapport à la CBIR. C'est pourquoi l'extraction d'autres types de caractéristiques est apparue avec les méthodes de CBVR afin d'utiliser le mouvement contenu dans la vidéo.

III.1.1.2.2 Caractéristiques dynamiques : Objets, Mouvement

Une première idée est de détecter et éventuellement reconnaître des **objets** dans les images et de les suivre au cours du temps. Dans le cas des vidéos médicales il s'agit par exemple de détecter les instruments médicaux. Lalys et al., par exemple, proposent de détecter les instruments de la chirurgie de la cataracte [21]. Ils cherchent tout d'abord à déterminer la présence du couteau qui possède un contour distinctif. La reconnaissance de l'instrument utilisé est une information forte, car elle est fortement corrélée à la réalisation des phases chirurgicales. Padoy et al. ont obtenu de très bons taux de reconnaissance en utilisant la présence des instruments pour segmenter temporellement leurs vidéos de chirurgies laparoscopiques en phases chirurgicales [25]. Cependant, cette étape de reconnaissance est complexe, car dans le cas de la chirurgie de la cataracte, la majorité des instruments a des formes similaires, et il est difficile de les différencier visuellement. Padoy et al. s'affranchissent de cette étape de reconnaissance et cette information est obtenue par un séquençage manuel réalisé par les chirurgiens [26]. Ils supposent que cette information pourra être connue dans le futur par l'utilisation de radio-étiquettes (puces RFID, « radio-frequency identification » en anglais) par exemple. La présence ou l'absence d'instruments chirurgicaux dans la scène chirurgicale, même sans les différencier, est déjà une information pertinente. Cao et al. cherchent, par exemple, à déterminer les instants d'entrée et de retrait des instruments pour identifier le début d'une phase d'action diagnostique ou thérapeutique [19].

Les **trajectoires** des objets peuvent également apporter des informations pertinentes. L'étude des trajectoires pour analyser une vidéo est très utilisée dans le domaine de la télésurveillance, notamment pour suivre des flux de voitures ou de piétons [44] [45]. Dans le domaine médical, les trajectoires des instruments chirurgicaux sont également souvent utilisées. Cette information est facilement disponible dans le cas de l'utilisation des robots chirurgicaux [31] [29]. Cependant Haro et al. ont montré que les méthodes basées sur la construction de sacs de mots visuels (paragraphe III.2.1.1.1) à partir de caractéristiques spatiotemporelles extraites des vidéos donnent des résultats équivalents, voire meilleurs que les méthodes de la littérature basées sur les trajectoires connues des outils chirurgicaux dans l'espace [29].

Il est également courant d'utiliser le **mouvement** global contenu dans la vidéo pour construire des signatures visuelles. Ces méthodes s'appuient généralement sur l'extraction du flux optique entre deux images consécutives. Pour cela, des points saillants sont détectés et sélectionnés puis le mouvement de ces points clés est calculé (flux optique). Quellec et al. utilisent le flux optique pour caractériser de courtes sous-séquences [49] et pour approcher un champ des vecteurs de déplacements au cours d'une courte séquence vidéo [62](paragraphe I.2.2.1, page 37). Droueche et al. utilisent la compression MPEG pour estimer les mouvements de blocs de pixels au sein de l'image et suivre des régions d'intérêts (paragraphe I.2.2.1) [47].

Ces caractéristiques visuelles sont ensuite utilisées pour construire les signatures visuelles des vidéos. Cela peut être fait de différentes manières, une des plus utilisées actuellement étant la construction de sac de mots visuels.

III.1.1.3 Construction des signatures visuelles

Il existe plusieurs manières d'utiliser ces caractéristiques pour construire une signature visuelle. Une première approche consiste à construire des **histogrammes** de répartition des couleurs, du mouvement (HOF pour « Histogram of Optical Flow » en anglais) ou des intensités des gradients horizontaux et verticaux (HOG). Cette approche est utilisée par Quellec et al. [49] ou Blum et al. [24] par exemple. Les histogrammes de répartition des couleurs sont aussi souvent utilisés pour construire les signatures visuelles. Cependant cette représentation ne permet pas de prendre en compte la répartition spatiale des couleurs. Lin et al. proposent alors trois alternatives de signatures basées sur la distribution spatiale des couleurs dans les images [63]. Les histogrammes ont l'avantage d'être simples à construire et sont donc peu coûteux en temps de calcul.

Les signatures sont généralement construites à partir d'une combinaison de différentes caractéristiques contenues dans les vidéos. Cela peut donner lieu à un vecteur de caractéristiques de grande taille. Or ce vecteur contiendra certainement une part d'information redondante ou non pertinente. Une approche pour pallier ce problème est de **réduire la dimension** du vecteur de caractéristiques, en conservant un maximum d'informations pertinentes. Cela est généralement réalisé par une analyse en composantes principales (ACP ou PCA - Principal Component Analysis -). Cette approche, utilisée par Gao et al. [64], ou Bashir et al. [65], permet de réaliser un changement de base de la base des variables initiale vers une base de variables non corrélées (composantes principales). Il existe d'autres méthodes de réduction de la dimension du vecteur de caractéristiques, comme l'analyse canonique des corrélations (ACC ou CCA -canonical-correlation analysis-) utilisée par Blum et al. [24]. Cette méthode permet de comparer deux groupes de variables pour savoir s'ils décrivent un même phénomène. Ainsi dans le cas de Blum et al., les

caractéristiques visuelles sont pondérées en se basant sur leurs corrélations avec l'utilisation des différents instruments chirurgicaux [24].

Une approche de construction de signatures très courante en CBIR et CBVR est l'utilisation de **sacs de mots visuels**. Il s'agit d'une approche issue de l'analyse de textes : un dictionnaire de mots est construit, puis un document est représenté par l'histogramme des mots qui le composent [66]. Dans le cas des images, les mots visuels sont construits à partir de descripteurs locaux, c'est-à-dire des caractéristiques visuelles extraites dans le voisinage de points d'intérêt. La méthode de construction des sacs de mots visuels est présentée dans la Figure 25.

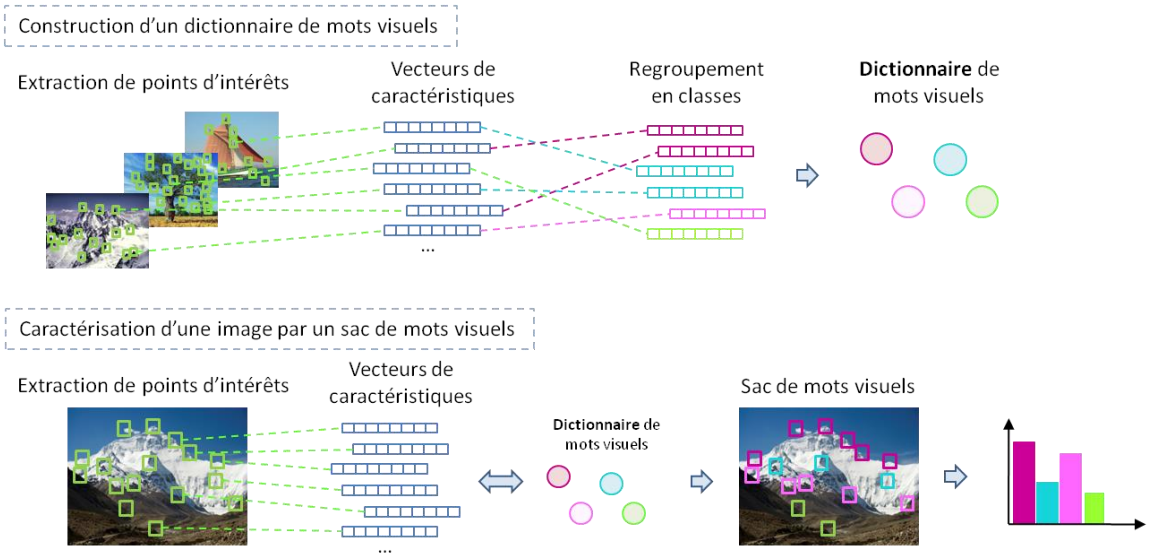


Figure 25. Méthode de caractérisation d'une image par sacs de mots visuels ; en haut, la construction d'un dictionnaire de mots visuels à partir de la base d'apprentissage ; en bas, la caractérisation d'une nouvelle image à partir d'un sac de mots visuels

Il existe plusieurs méthodes de construction de descripteurs locaux. L'une des plus connues est le calcul de descripteurs SIFT (Scale-Invariant Feature Transform, transformation de caractéristiques visuelles invariante à l'échelle) de Lowe [67]. André et al. utilisent un détecteur dense, adapté à l'endomicroscopie, fait de disques qui se chevauchent, combiné avec un descripteur SIFT [56] [32]. Il existe d'autres types de descripteurs, tels que SURF (Speeded Up Robust Features : caractéristiques robustes accélérées). Le détecteur de points d'intérêt de SURF est invariant aux changements d'échelle et aux rotations. Dans le cas des vidéos, la notion de points d'intérêt est étendue au domaine spatiotemporel. Les STIP (Space-Time Interest Points) sont des points d'intérêt spatiotemporels proposés par Laptev et al. [68] et largement utilisés dans la littérature. Ils sont obtenus grâce à un détecteur de Harris [69] étendu dans le domaine temporel, appliqué à différentes échelles spatiales et temporelles. Ils représentent ainsi des points de variations importantes à la fois dans le domaine spatial et dans le domaine temporel. Les caractéristiques visuelles sont alors extraites dans des volumes spatiotemporels autour des STIP. De leur côté, Haro et al. [29], Zappella et al. [31], extraient des informations locales (caractéristiques HOG et HOF) contenues dans des parallélépipèdes rectangles centrés sur les points d'intérêt. Ce type de caractérisation en sacs de mots visuels permet d'obtenir de très bons résultats [31]. Cependant cette méthode est coûteuse en temps de calcul et n'est pas compatible avec une utilisation en temps réel.

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

Il existe de nombreuses autres manières de construire les signatures de vidéo. Lalys et al., par exemple, construisent leurs vecteurs de caractéristiques à partir de la sortie de différents classifieurs [21]. Cela permet d'avoir une signature très complète, très proche de la description sémantique de la chirurgie de la cataracte, mais coûteuse également en temps de calcul.

Une fois la bonne signature visuelle choisie pour l'application voulue, cette signature va permettre de comparer deux vidéos, ou deux segments de vidéos, entre eux. Pour cela il est nécessaire de choisir une mesure de similitude.

III.1.2 Mesure de similitude

III.1.2.1 Alignement dynamique temporel (DTW)

La réalisation des chirurgies par des praticiens différents, pour des patients différents ajoute une contrainte supplémentaire à la mesure de similitude. En effet, les vidéos sont généralement de durées différentes et elles représentent des actions dont les vitesses d'exécution sont également variables. L'algorithme **DTW** (alignement dynamique temporel ou « Dynamic Time Warping » en anglais) introduit par Berndt et al. [70] apporte une réponse à ce problème. Le principe de l'algorithme est présenté dans la Figure 26. DTW recherche un appariement optimal entre deux séries temporelles $\mathcal{X} = \{x_1, x_2, \dots, x_m\}$ et $\mathcal{Y} = \{y_1, y_2, \dots, y_n\}$ de tailles respectives m et n. Pour cela, on cherche le chemin de moindre coût dans une matrice de mesure de coût locale $\boldsymbol{D} = \left(d_{ij}\right)_{0<i<m, 0<j<n}$ où d_{ij} représente la mesure (généralement une distance euclidienne) entre x_i et y_j. Le chemin w est déterminé en allant de $w(m,n)$ vers $w(0,0)$ et la distance minimale entre les deux séquences est donnée par la somme des éléments du chemin w. Il existe différentes variantes de l'algorithme DTW, telles que sparseDTW [71], ou FDTW qui limitent la recherche par l'utilisation d'une enveloppe [72].

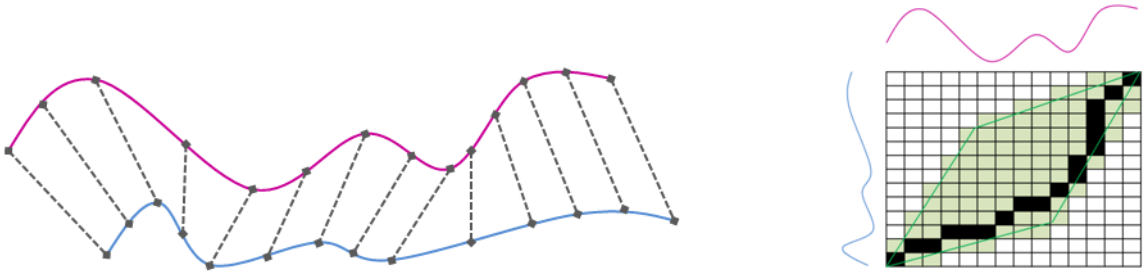


Figure 26. Principe de l'algorithme DTW ; à gauche, l'alignement de deux séquences réalisé avec l'algorithme DTW ; à droite, le chemin de moindre coût avec en vert un exemple d'enveloppe de restriction de la recherche

La mesure de distance entre deux séquences via l'algorithme DTW a l'avantage de gérer les différences de durées et de vitesses d'exécution des actions, mais elle nécessite de connaître l'intégralité de la séquence requête pour être calculée. Cette approche n'est donc pas compatible avec une utilisation « en direct ».

III.1.2.2 Mesure de similitude de Piciarelli et al.

Piciarelli et al. ont proposé une mesure alternative, permettant de calculer la mesure de similitude de façon progressive : la mesure est mise à jour à chaque acquisition d'une nouvelle image [44]. Cette mesure de similitude a été développée dans le cadre de l'analyse automatique de trajectoires de véhicules. Comme cela est présenté dans la Figure 27, l'ensemble des trajectoires de la base d'apprentissage est partitionné et modélisé par un ensemble de nœuds organisés selon une structure en arbre.

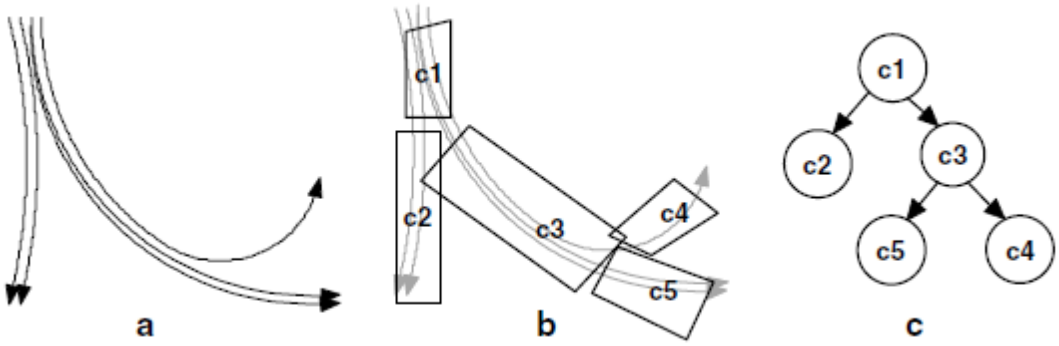


Figure 27. Partitionnement de trajectoires selon la méthode de Piciarelli et al. [44]

Chaque trajectoire $\mathcal{T}$ est définie par un ensemble de positions $\{t_1, t_2, \dots, t_n\}$ relevées à intervalles réguliers où $t_i = \{x_i, y_i\}$. Chaque nœud $\mathcal{C}$ représente une section de trajectoire moyenne $\{c_1, c_2, \dots, c_m\}$, où $c_j = \{x_j, y_j, \sigma_j^2\}$ avec σ_j^2 une approximation de la variance locale à la position j. Lors de la construction de l'arbre, pour chaque nouvelle trajectoire, la similitude entre la trajectoire en cours et chaque nœud de l'arbre est calculée à l'aide d'une fenêtre glissante (Figure 28). Si une correspondance est trouvée, alors le nœud est mis à jour. Si le point $c_j = \{x_j, y_j, \sigma_j^2\}$ est l'élément le plus proche du nouveau point $t_i = \{x_i, y_i\}$ alors le point est mis à jour de la manière suivante :

$$\begin{cases} x_j = (1 - \alpha)x + \alpha x_i \\ y_j = (1 - \alpha)y + \alpha y_i \\ \sigma_j^2 = (1 - \alpha)\sigma_j^2 + \alpha d_{ij}^2 \end{cases}$$

où α est un réel entre 0 et 1. Quand la distance entre la trajectoire et le nœud devient supérieure à un seuil, la trajectoire sort du nœud. Cela amène à une nouvelle étape de mise en correspondance et éventuellement à une division du nœud (dans le cas où la trajectoire sort du nœud sans être proche de la fin de celui-ci).

La similitude entre une trajectoire $\mathcal{T} = \{t_1 \dots t_n\}$ et un groupe de trajectoires $\mathcal{C} = \{c_1 \dots c_m\}$ est calculée de la manière suivante :

$$D(\mathcal{T}, \mathcal{C}) = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} d(t_i, \mathcal{C})$$

où

$$d(t_i, \mathcal{C}) = \min_j \left(\frac{d_{ij}}{\sqrt{\sigma_j^2}} \right), j \in \{[(1-\delta)i] \dots [(1+\delta)i]\}$$

et d_{ij} est la distance euclidienne entre un point t_i de la trajectoire $\mathcal{T}$ et un point c_j du groupe de trajectoires $\mathcal{C}$. La distance entre une trajectoire et un nœud est alors la moyenne des distances normalisées entre chaque point t_i de la trajectoire $\mathcal{T}$ et le point le plus proche dans la trajectoire moyenne du groupe $\mathcal{C}$ au sein d'une fenêtre glissante centrée sur i. La taille de la fenêtre glissante dépend de la position i dans la trajectoire et grandit au cours du temps. Le paramètre δ représente ce facteur d'agrandissement. Le principe de l'algorithme de mesure de similitude est présenté dans la Figure 28.

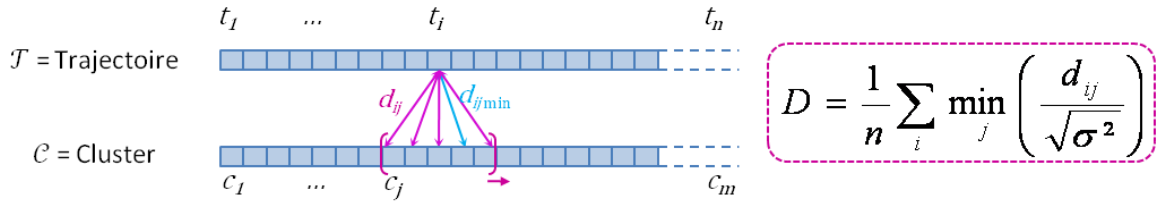


Figure 28. Principe de la mesure de similitude proposée par Piciarelli et al. [44]

Nous pouvons imaginer utiliser cette méthode pour comparer deux vidéos en temps réel. Cette méthode a l'avantage d'être simple à implémenter et calcule la mesure de similitude tout au long de l'acquisition de la vidéo, sans nécessiter d'attendre la fin de l'enregistrement. De plus, l'utilisation de la structure en arbre peut être une alternative aux arbres k-d (paragraphe III.1.3) en permettant de restreindre l'espace de recherche. Ce point est particulièrement intéressant du fait de notre contrainte de temps réel. Nous pouvons également envisager d'utiliser par la suite cette méthode pour construire un modèle du déroulement de la chirurgie de la cataracte, et faciliter ainsi l'analyse d'une vidéo de chirurgie complète.

III.1.3 Catégorisation des vidéos

Grâce à une mesure de similitude adaptée, les cas les plus proches peuvent être trouvés au sein de la base de données par une recherche des K plus proches voisins. On cherche donc les cas dans la base de données dont les signatures visuelles sont les plus proches de la signature du cas placé en requête. Il existe plusieurs algorithmes permettant de réaliser cette recherche.

Le premier algorithme, la **recherche linéaire**, est le plus simple. Il consiste à parcourir l'ensemble des cas de la base et à regarder si ce point est plus proche ou non qu'un des plus proches voisins déjà sélectionnés. Cette méthode est simple et intuitive, cependant, elle peut s'avérer coûteuse en temps de calcul lorsque le nombre de cas archivés dans la base augmente.

Un autre type d'algorithmes de recherche, plus efficaces, consiste à partitionner l'espace de recherche, en utilisant un **arbre k-d** par exemple. Cela permet de structurer la recherche. Cette méthode est utilisée par exemple par Gao et al. [73] ou Vlachos et al. [74]. Cependant, lorsque la

dimension D de l'espace des paramètres devient trop grande, cette méthode peut également s'avérer coûteuse en temps de calcul.

Il existe des méthodes très rapides de recherche, dites « **Local Sensitive Hashing** », qui se basent sur des tables de hachages. Chaque signature représentant un cas est remplacée par une clé calculée par une fonction de hachage. La recherche des plus proches voisins se fait alors par comparaison des clés. Il s'agit d'une méthode approximative, mais qui permet d'accélérer considérablement les recherches. Cette méthode est utilisée dans [52], [53] ou [54].

III.1.4 Synthèse

L'étude des différentes méthodes de caractérisation et de comparaison des vidéos nous a permis de dégager et de faire un premier choix des méthodes adaptées à notre problème de reconnaissance automatique des tâches chirurgicales. La littérature propose plusieurs réponses à nos objectifs et nos contraintes. Tout d'abord, il semble ressortir que l'étude du mouvement dans la vidéo est un élément à prendre en compte dans la caractérisation des vidéos. Les signatures sous forme d'histogrammes de mots visuels (sacs de mots visuels) sont fréquemment utilisées et semblent fournir des résultats satisfaisants. Cette méthode semble en revanche un peu coûteuse en calcul, et n'est peut-être pas adaptée à notre contrainte de temps réel. Il s'agit néanmoins d'une bonne méthode de référence pour évaluer nos méthodes. Il semble également intéressant de détecter et de suivre les outils chirurgicaux. Cependant, cette tâche n'est pas aisée du fait de la grande ressemblance entre les outils dans notre cas. Ce choix rend de plus la méthode très spécifique à la chirurgie, et plus particulièrement à la chirurgie de la cataracte, contrairement à l'utilisation de signatures plus générales, basées sur le mouvement ou la couleur par exemple. En revanche, dans le cas de la chirurgie de la cataracte, la seule analyse des couleurs des images ne peut être suffisante pour caractériser nos différentes tâches chirurgicales. En effet il y a peu de variations de couleur tout au long de la chirurgie. La couleur des structures anatomiques évolue peu et les différents outils utilisés ont dans la majorité des cas une couleur grise. Enfin, la mesure de similitude proposée par [44] semble être une bonne alternative à la mesure de distance DTW. Elle est compatible avec une utilisation « en direct » et est simple à mettre en place. Plus généralement la méthode de construction d'arbre proposée par Piciarelli et al. [44] offre également des possibilités de recherches rapides (alternative aux kd-tree) et de modélisation du processus chirurgical. Les méthodes choisies et mises en place sont présentées dans les sections III.2 et III.3 suivantes.

III.2 Caractérisation des vidéos : nos choix

La première étape de notre méthode de reconnaissance automatique de tâches chirurgicales consiste à caractériser les séquences vidéo en les représentant par une signature visuelle. Pour cela nous avons décidé de nous appuyer sur l'extraction de caractéristiques basées sur le mouvement. En effet, d'après les informations de la littérature, ce type de caractéristiques semble intéressant et particulièrement adapté à l'analyse des vidéos. Nous avons donc choisi d'étudier deux méthodes d'extraction de caractéristiques basées sur l'extraction du mouvement dans la vidéo. Nous avons également évalué une méthode de normalisation des vidéos visant à améliorer la pertinence des caractéristiques visuelles extraites, en compensant par exemple les mouvements parasites.

III.2.1.1 Extraction de caractéristiques basées sur le mouvement

Nous cherchons à apporter une aide en temps réel au chirurgien en comparant une vidéo en cours d'acquisition à des vidéos archivées de la même chirurgie. Nous avons vu précédemment (paragraphe III.1.1.2.2) que l'étude du mouvement était une bonne méthode pour caractériser le contenu d'une vidéo dans ce cadre. Nous avons choisi d'évaluer deux méthodes de construction de signatures visuelles utilisées dans la littérature, toutes deux basées sur l'extraction du mouvement dans la vidéo.

III.2.1.1.1 Histogrammes de mots visuels

Le premier type de signatures visuelles utilisé est basé sur la construction de sacs de mots visuels. Il s'agit d'une méthode de référence de la littérature. Le principe de la construction des signatures à base de sacs de mots visuels est présenté dans la Figure 25. Des caractéristiques locales (descripteurs) sont extraites dans les vidéos de la base de cas puis regroupées en différentes classes : les « mots visuels ». Ces mots visuels constituent un dictionnaire à partir duquel sont construites les signatures visuelles. Chaque descripteur local est associé à un mot du dictionnaire, puis un histogramme de répartition des mots visuels est construit. Cet histogramme, dit « histogramme de mots visuels » (« Visual Word Histogram » en anglais), constitue la signature de la vidéo, ou d'un segment de la vidéo. Dans notre cas nous construisons un histogramme de mots visuels par image, et l'ensemble de ces histogrammes de mots visuels constitue la signature de notre vidéo.

Les descripteurs utilisés sont extraits dans le voisinage de points d'intérêt spatiotemporels, les STIP proposés par Laptev [68]. Le détecteur de points d'intérêt STIP peut être vu comme une extension dans le domaine spatiotemporel d'un détecteur d'angles (détecteur de Harris). Mais les STIP sont également détectés en correspondance avec le mouvement, et les informations contenues dans les zones statiques de l'image sont automatiquement éliminées du descripteur. Un exemple de points d'intérêts spatiotemporels STIP extraits dans deux images de vidéos de chirurgies de la cataracte est présenté dans la Figure 29. Les caractéristiques visuelles locales sont ensuite extraites à l'intérieur de parallélépipèdes rectangles centrés sur les points saillants trouvés par le détecteur. Ces caractéristiques sont utilisées pour construire un histogramme à 72 classes des amplitudes des gradients horizontaux et verticaux (HOG) et un histogramme à 90 classes du

flux optique (HOF). L'extraction de ces descripteurs dans le voisinage de points d'intérêt STIP est réalisée à l'aide de l'exécutable fourni par Ivan Laptev[4].

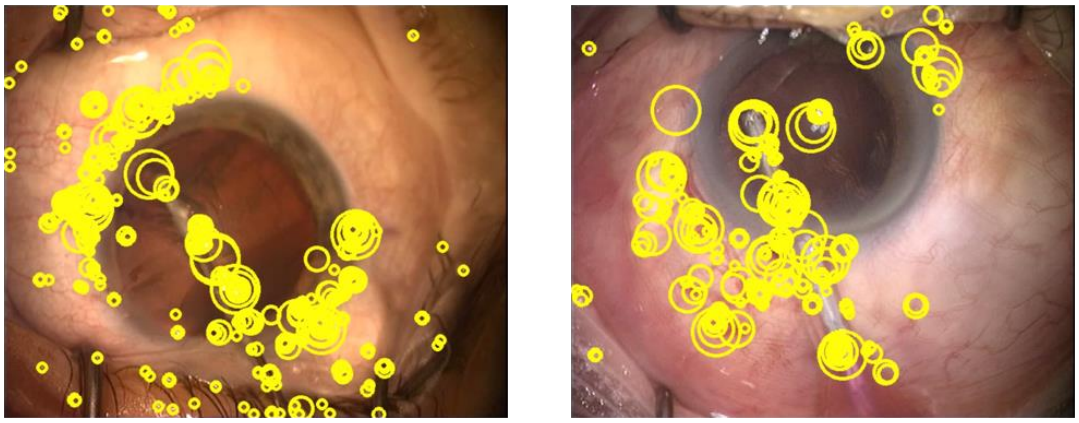


Figure 29. Exemple de points d'intérêts spatiotemporels STIP extraits dans deux images de vidéos de chirurgies de la cataracte

Ce type de signatures a prouvé son efficacité dans de nombreuses méthodes, dont des méthodes développées pour la chirurgie [29][31]. Elle est cependant coûteuse en temps de calcul et n'est donc pas compatible avec notre contrainte de temps réel. Cette méthode nous servira donc de référence pour tester nos algorithmes et un autre type de signatures visuelles sera testé : moins coûteux en temps de calcul, il est basé sur l'extraction du flux optique.

III.2.1.1.2 Histogrammes de Mouvement

Une seconde méthode de construction de signatures visuelles, basée sur l'extraction du flux optique entre deux images consécutives de la vidéo, a été évaluée. Cette méthode a été utilisée par Quellec et al. [49] et consiste en la construction d'histogrammes de mouvements.

Les caractéristiques de mouvement de la vidéo sont extraites en calculant le flux optique entre deux images consécutives $\boldsymbol{I}_{i-1}$ et $\boldsymbol{I}_i$. Des points saillants (angles) sont extraits au sein des images en les sélectionnant parmi l'ensemble des pixels p de l'image en fonction de la plus petite valeur de la matrice $\boldsymbol{M}_p$ si ci-dessous :

[4] http://www.di.ens.fr/~laptev/interestpoints.html

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

$$\begin{cases} \boldsymbol{M}_p = \begin{pmatrix} \boldsymbol{A}_p & \boldsymbol{B}_p \\ \boldsymbol{B}_p & \boldsymbol{C}_p \end{pmatrix} \\ \boldsymbol{A}_p = \sum_{(x,y)\in\mathcal{N}_p} \left(\frac{\partial \boldsymbol{I}_i}{\partial x}(x,y)\right)^2 \\ \boldsymbol{B}_p = \sum_{(x,y)\in\mathcal{N}_p} \frac{\partial \boldsymbol{I}_i}{\partial x}(x,y).\frac{\partial \boldsymbol{I}_i}{\partial y}(x,y) \\ \boldsymbol{C}_p = \sum_{(x,y)\in N_p} \left(\frac{\partial \boldsymbol{I}_i}{\partial y}(x,y)\right)^2 \end{cases}$$

où $\mathcal{N}_p$ est un voisinage du pixel p. Le flux optique entre $\boldsymbol{I}_{i-1}$ et $\boldsymbol{I}_i$ est ensuite calculé entre chaque point saillant en s'appuyant sur la méthode itérative de Lucas-Kanade [75]. La librairie OpenCV[5] est utilisée pour extraire les points saillants et calculer le flux optique. Quatre histogrammes de huit classes sont alors calculés pour caractériser le mouvement. Le premier histogramme représente l'amplitude des vecteurs de mouvement. Les deux suivants représentent les coordonnées (un pour les coordonnées x et un pour les coordonnées y) et enfin, le dernier représente les directions. Un vecteur de 32 caractéristiques est donc construit pour chaque paire d'images consécutives.

Cette méthode permet de représenter le mouvement dans la vidéo et est peu coûteuse en temps de calcul. Cependant, il existe des mouvements parasites dans les vidéos, car, contrairement aux images de vidéosurveillance, l'arrière-plan n'est pas fixe. C'est pourquoi nous avons du mettre en place une étape de recalage et normalisation spatiale des vidéos.

III.2.1.2 Normalisation des vidéos

Les deux méthodes de construction de signatures visuelles présentées dans le paragraphe précédent (paragraphe III.2.1.1) sont très génériques et peuvent être appliquées à tous types de vidéos. Cela a l'avantage de rendre nos algorithmes indépendants du type de vidéos utilisées, et de pouvoir les appliquer facilement à d'autres types de chirurgies que la chirurgie de la cataracte. Nous avons néanmoins décidé de réaliser un prétraitement propre aux chirurgies du segment antérieur, pour affiner l'extraction de nos caractéristiques visuelles. Il existe différentes sources de mouvements au sein des vidéos de chirurgies de la cataracte. La principale est le mouvement des instruments dans le champ de vue de la caméra. C'est le mouvement le plus pertinent, car directement lié au geste chirurgical pratiqué. Il y a également les mouvements de l'œil. Ceux-ci peuvent être liés au fait que le patient est le plus souvent sous anesthésie locale, et peut donc bouger son œil. Ils peuvent également être induits par l'action des instruments chirurgicaux sur le globe oculaire. Ces mouvements peuvent alors s'avérer non pertinents pour caractériser les vidéos et reconnaître le geste chirurgical. Enfin, des mouvements de la caméra peuvent parasiter la caractérisation des vidéos. Le facteur de zoom est également variable d'une chirurgie à l'autre ou au sein d'une même chirurgie. Cela rend la comparaison des signatures complexe, en introduisant des mouvements parasites. De plus, la zone d'action est sans cesse déplacée et est de taille variable.

5 http://opencv.org/

Enfin, comme cela est montré dans la Figure 30, les gestes chirurgicaux ont lieu dans une zone centrée sur la pupille et l'iris. Il n'est pas nécessaire d'extraire des caractéristiques visuelles en dehors de cette zone, sur les paupières par exemple.

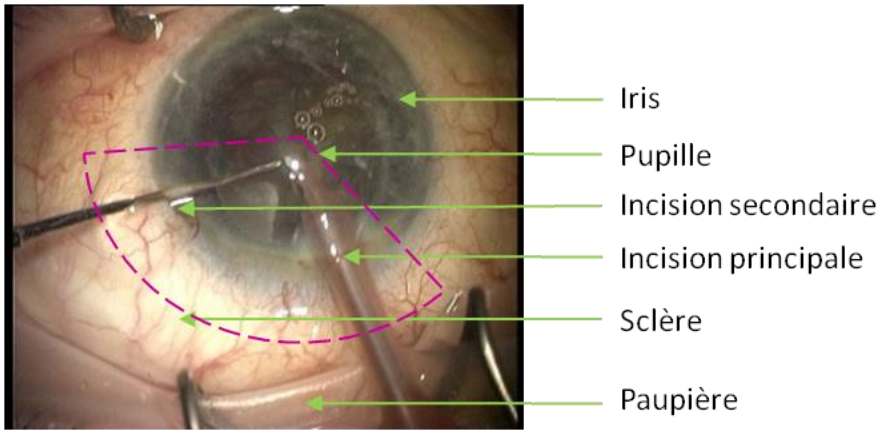


Figure 30. Différentes zones de l'œil visualisées dans le champ de vue de la caméra ; en rose, la zone d'action principale des outils

A partir de ces constatations, nous avons alors mis en place trois types de normalisations basés sur la détection du centre de l'iris et de la pupille et sur l'estimation du facteur d'échelle [76]. La détection du centre de l'iris et de la pupille et l'estimation du facteur de zoom sont réalisées sans explicitement segmenter la pupille ou l'iris. Dans un premier temps, le centre est détecté en s'appuyant sur la transformée de Hough. Les centres sont détectés grâce à un accumulateur 2D. Les cercles formés par la pupille et l'iris étant approximativement concentriques, les informations de leurs contours s'accumulent dans la même région de l'accumulateur. En revanche il est parfois difficile de distinguer les frontières exactes entre la pupille et l'iris et entre l'iris et la sclère (Figure 30). Cela est dû aux variations importantes de textures et de couleurs de l'iris, de la sclère ou des paupières, d'un patient à l'autre. Nous proposons d'estimer le facteur d'échelle en détectant les positions des reflets de l'éclairage du microscope sur la cornée. Ces reflets forment trois taches lumineuses, elles sont visibles dans les images de la Figure 31. Ces reflets sont présents dans la quasi-totalité des images et le motif dépend uniquement de la forme de la cornée et de la distance entre l'œil et l'éclairage. Or la forme de la cornée est très peu variable d'un patient à l'autre. Ainsi la taille du motif des reflets est majoritairement contrôlée par le facteur de zoom, et cela de façon linéaire.

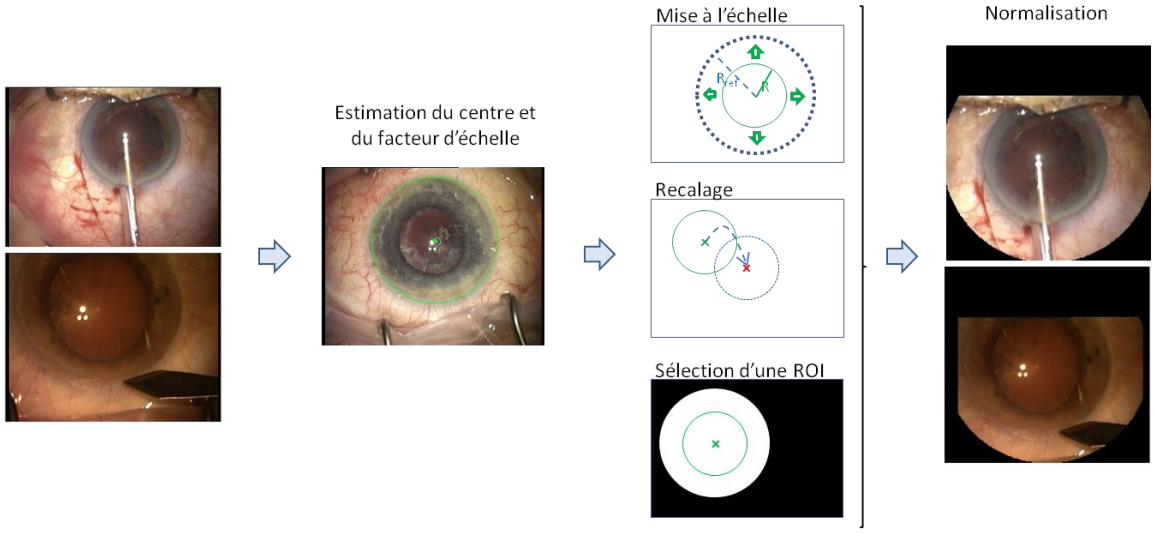


Figure 31. Normalisation des vidéos de chirurgie de la cataracte à partir de la détection du centre (de la pupille et de l'iris) et du facteur d'échelle

Le principe de la normalisation des vidéos à partir du centre et du facteur d'échelle est présenté dans la Figure 31. Trois normalisations sont appliquées aux images de la vidéo pour les normaliser spatialement : la mise à l'échelle, le recalage et la sélection d'une région d'intérêt (ROI pour « Region Of Interest » en anglais). La **mise à l'échelle** des vidéos permet de compenser les variations dues aux changements de zoom, en normalisant les images pour avoir un facteur de zoom identique dans chacune d'entre elles. Le **recalage** permet de compenser les mouvements parasites, tels que les mouvements de l'œil et de la caméra en centrant toutes les images des vidéos sur le centre de l'iris. Enfin, la sélection d'une région d'intérêt (**ROI**) permet d'extraire des signatures visuelles en se limitant à une zone pertinente de l'œil. Cette zone correspond à la zone d'action des outils, c'est-à-dire un cercle centré sur le centre de l'iris. Un masque circulaire centré sur l'iris et la pupille, avec un rayon légèrement plus grand que le rayon de l'iris, a été appliqué sur les images.

Cette étape de normalisation améliore la comparaison des images, l'extraction des informations pertinentes, notamment pour les mouvements des outils, et permet donc de mieux reconnaître des gestes qui sont semblables.

III.3 Catégorisation des vidéos : nos choix

La catégorisation des vidéos va nous permettre de déterminer le type des tâches effectuées par comparaison de la vidéo requête aux cas de la base de données, en y cherchant ses plus proches voisins. Le résultat de cette recherche permet de déterminer la tâche la plus probablement représentée dans la séquence. Nous avons choisi de nous inspirer de la mesure de similitude proposée par Piciarelli et al. [44] pour comparer les signatures visuelles, et d'évaluer ses performances dans le domaine de la chirurgie.

III.3.1 Mesure de similitude de Piciarelli et al.

Nous avons vu au paragraphe III.1.2.2 que la mesure de similitude de Piciarelli et al. a l'avantage de ne pas nécessiter d'attendre d'avoir enregistré l'intégralité de la séquence requête pour être calculée [44]. La mesure de similitude est calculée de façon progressive, c'est-à-dire qu'elle est mise à jour à chaque acquisition d'une nouvelle image. Cette mesure de similitude a été développée initialement pour comparer la trajectoire d'un véhicule en déplacement avec une trajectoire moyenne. A chaque nouveau point de la trajectoire enregistré, ce dernier est comparé aux points de la trajectoire moyenne contenus au sein d'une fenêtre glissante. Cette fenêtre est centrée, dans la trajectoire moyenne, sur le point de même indice i que le nouveau point acquis, et elle grandit proportionnellement à cet indice.

Dans notre cas, nous remplaçons les trajectoires par des séquences vidéo. Par analogie, les points de la trajectoire sont maintenant les vecteurs de caractéristiques des images des séquences vidéo. Nous cherchons à trouver les plus proches voisins au sein de la base de cas archivés, la vidéo en cours d'acquisition n'est donc pas comparée à une vidéo moyenne, mais à tous les cas de la base de données. La mesure de distance utilisée entre la vidéo requête $\mathcal{V}_1 = \{v_{10}, \dots, v_{1n}\}$ et une vidéo de la base de cas $\mathcal{V}_2 = \{v_{20}, \dots, v_{2m}\}$ est alors définie de la manière suivante :

$$D(\mathcal{V}_1, \mathcal{V}_2) = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} d(v_{1i}, \mathcal{V}_2)$$

où

$$d(v_{1i}, \mathcal{V}_2) = \min_j \left(d_{ij}\right), j \in \{[(1-\delta)i] \dots [(1+\delta)i]\}$$

La mesure de distance locale d_{ij} entre les éléments x_i et y_j des séries temporelles $\mathcal{X}$ (référence) et $\mathcal{Y}$ (requête) utilisée par Piciarelli et al. est généralement une distance **euclidienne** définie dans l'équation suivante :

$$d_{ij} = \sqrt{\sum_{k=1}^{l} (x_{ik} - y_{jk})^2}$$

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

Cependant les caractéristiques visuelles extraites à chaque pas de temps sont dans notre cas sous la forme d'un vecteur de caractéristiques représentant les valeurs d'un histogramme. D'autres mesures locales peuvent être mieux adaptées à la comparaison d'histogrammes, comme les trois mesures suivantes. La première est la mesure **Chi-Square**, qui est définie de la manière suivante :

$$d_{ij} = \sum_{k=1}^{l} \frac{(x_{ik} - y_{jk})^2}{x_{ik}}$$

Une seconde est la distance de **Bhattacharyya**, définie de la manière suivante :

$$d_{ij} = \sqrt{1 - \frac{1}{\sqrt{\bar{x}_i \bar{y}_j l}} \sum_{k=1}^{l} (x_{ik} \cdot y_{jk})}$$

Enfin, la dernière est la **corrélation**, définie de la manière suivante :

$$d_{ij} = \frac{\sum_{k=1}^{l} (x_{ik} - \bar{x}_i)(y_{jk} - \bar{y}_j)}{\sqrt{\sum_{k=1}^{l} (x_{ik} - \bar{x}_i)^2 \sum_{k=1}^{l} (y_{jk} - \bar{y}_j)^2}}$$

Les différentes mesures de distance locale d_{ij} ont été évaluées, pour choisir la mesure la plus adaptée à notre problème. La mesure de distance entre deux vidéos va ensuite permettre de chercher les cas les plus proches au sein de la base de données.

III.3.2 Recherche des plus proches voisins

Les K plus proches voisins de la vidéo requête sont recherchés en s'appuyant sur la mesure de similitude présentée dans le paragraphe précédent (paragraphe III.3.1). La taille de la base étant raisonnable (paragraphe III.4.1), la recherche se fait de façon linéaire en parcourant l'ensemble des cas de la base. Néanmoins, dans le cas de l'utilisation d'une base de données plus conséquente, on pourra envisager d'utiliser des méthodes de recherche plus rapides en s'appuyant par exemple sur des kd-trees ou des tables de hachage ou en s'appuyant sur les trajectoires moyennes définies au paragraphe III.1.2.2. Le principe de la méthode est présenté dans la Figure 32. Les tâches représentées par les K plus proches voisins permettent de calculer une probabilité d'appartenance de la séquence requête aux différentes tâches chirurgicales.

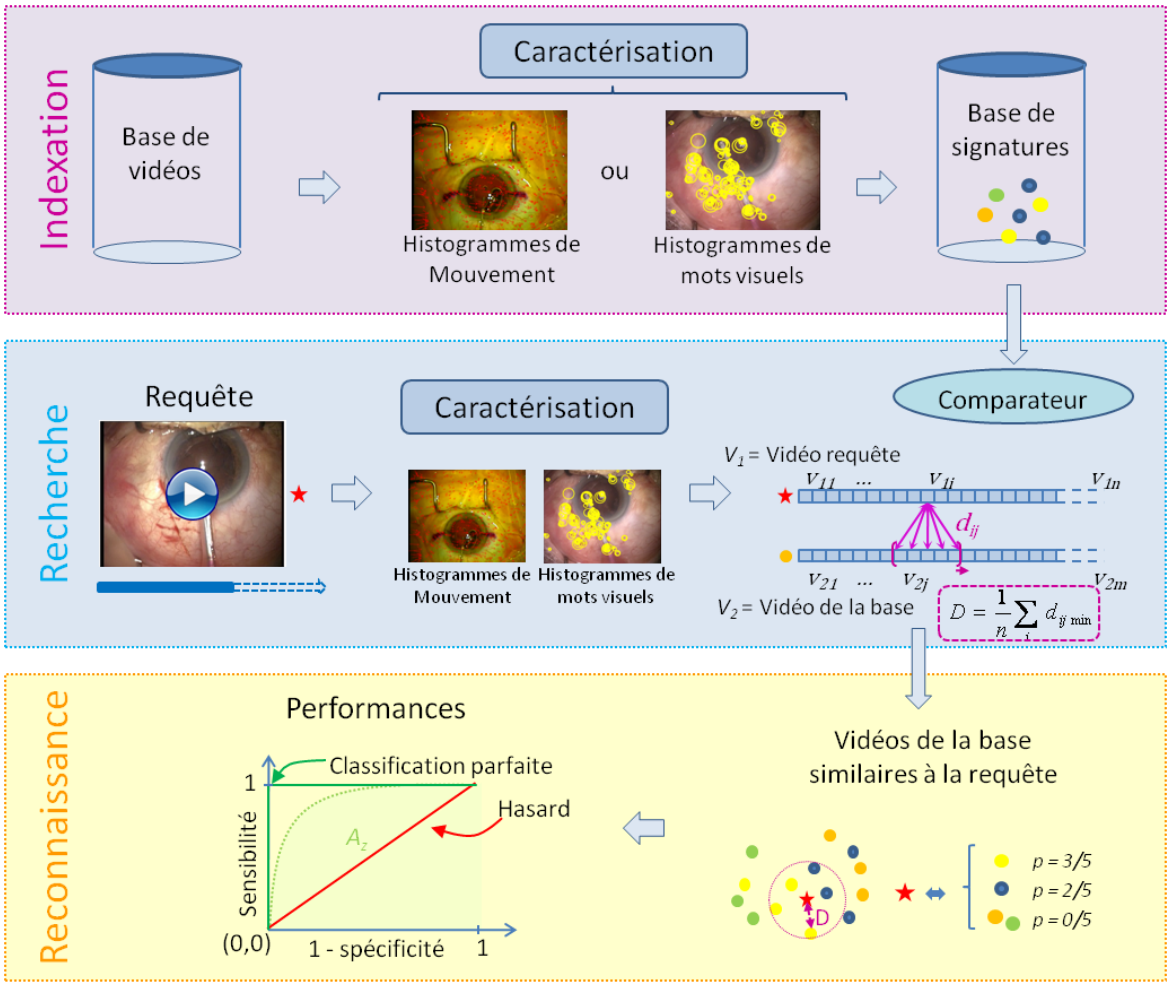


Figure 32. Principe de la méthode de reconnaissance automatique des tâches chirurgicales

Les performances de reconnaissance de la méthode ont ensuite été évaluées en termes d'aire sous la courbe ROC. Les différents points de la méthode ont été évalués et les résultats sont présentés et discutés dans la partie suivante (paragraphe III.4).

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

III.4 Evaluation

La méthode de reconnaissance automatique de tâches chirurgicales mise en place a été évaluée sur une base de séquences vidéo où une séquence représente une tâche chirurgicale. Les performances ont été évaluées pour les différentes variations de la mesure de similitude, les différentes normalisations et les deux types de signature visuelles.

III.4.1 La base de sous-séquences vidéos

Les méthodes de reconnaissance développées sont évaluées sur une sous-base de 30 séquences vidéo sélectionnées parmi la base de données initiale présentée dans le paragraphe II.2.2 . Nous avons décidé en effet de n'utiliser que 30 vidéos, par souci de comparaison avec le Chapitre IV, dans lequel nous évaluons une méthode basée sur la nouvelle description multi-échelles. Cette sous-base, représentative de la base de données principale, a été constituée par David Martiano, interne en chirurgie au CHRU de Brest, et moi-même, afin d'être réinterprétée selon une description multi-échelles de la chirurgie (paragraphe II.3). Pour l'évaluation et le choix des paramètres, une autre sous-base de 30 séquences vidéos extraites de la base initiale a été utilisée (base d'apprentissage). Nous avons cherché à reconnaître automatiquement dans les vidéos de la base de test les neuf tâches chirurgicales principales qui composent la chirurgie de la cataracte. Les 60 vidéos ont été découpées en séquences vidéo où chacune des sous-séquences représente une tâche chirurgicale. Au total, 303 séquences ont été obtenues pour la base de test et 304 séquences ont été obtenues pour la base d'apprentissage. Les statistiques de représentation des tâches chirurgicales au sein de la base de test et leurs durées moyennes sont présentées dans le Tableau 3.

Tableau 3. Statistiques de la base de séquences vidéo utilisées pour évaluer la méthode de reconnaissance automatique de tâches chirurgicales (base de test)

Tâches chirurgicales	nombre de séquences	durée moyenne (+/- écart type)
Incision	32	1:07 (+/- 0:17)
Rhexis	30	1:30 (+/- 0:14)
Hydrodissection	29	0:33 (+/- 0:04)
Phacoémulsification	30	3:09 (+/- 0:15)
Epinoyau	30	1:40 (+/- 0:10)
Injection visqueux	62	0:14 (+/- 0:02)
Mise en Place	30	0:39 (+/- 0:04)
Retrait Visqueux	30	0:57 (+/- 0:07)
Fermeture	30	2:16 (+/- 0:14)

III.4.2 Mesure de la performance : l'aire sous la courbe ROC

Les performances ont été évaluées en termes d'aire sous la courbe ROC. La courbe ROC, également appelée courbe sensibilité/spécificité, permet d'évaluer les performances d'un classifieur binaire. La courbe représente le taux de vrais positifs en fonction du taux de faux positifs. Un classifieur binaire réalise un choix entre deux possibilités (positif et négatif). Un vrai positif (VP) est alors un élément positif correctement détecté par le classifieur et un faux positif (FP) un élément négatif déclaré positif. Par exemple dans le cas d'une classification pathologique/non pathologique les possibilités de résultats sont présentées dans le tableau suivant :

Tableau 4. Possibilités de résultats d'un classifieur binaire pathologique/non pathologique

	pathologique	Non pathologique
Test positif	VP	FP
Test négatif	FN	VN

Les faux négatifs (FN) représentent les cas pathologiques déclarés négatifs et les vrais négatifs (VN) représentent les cas non pathologiques déclarés négatifs. Dans notre cas, les cas reconnus pathologiques sont les séquences bien classées, les non pathologiques les séquences mal classées. Comme présenté dans la Figure 33, la courbe ROC représente le taux de vrais positifs (sensibilité) :

$$sensibilité = \frac{VP}{VP + FN}$$

en fonction du taux de faux positifs (un moins la spécificité) :

$$1 - spécificité = 1 - \frac{VN}{VN + FP}$$

Le principe de lecture de la courbe ROC est présenté dans la Figure 33, dans laquelle la courbe verte représente une classification parfaite et la courbe rouge une classification aléatoire.

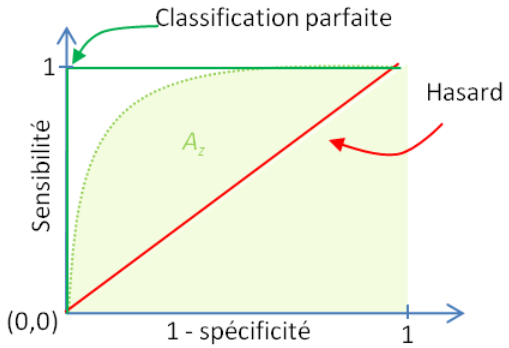


Figure 33. Présentation de la courbe ROC

Au point $(0, 0)$ de la courbe, le classifieur déclare tous les éléments négatifs : il n'y a aucun faux positif, tous les cas sont déclarés non pathologiques. Au point $(1, 1)$ le classifieur déclare tous les éléments positifs, aucun cas pathologique n'est déclaré sain, mais le classifieur ne différencie rien. Au point $(1,0)$, il n'y a aucun vrai positif et aucun vrai négatif, les cas pathologiques sont

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

déclarés sains et les cas non pathologiques sont déclarés pathologiques, le classifieur se trompe toujours. Un bon classifieur doit se rapprocher du point $(1,1)$ qui représente une classification parfaite. En effet, l'idéal est d'avoir une très bonne sensibilité (très peu de faux négatifs) afin de ne rater aucun cas pathologique, tout en étant spécifique (peu de faux positifs). L'aire A_z sous la courbe ROC est donc généralement comprise entre $0{,}5$ (classifieur aléatoire) et 1 (classifieur parfait) et l'objectif est d'avoir une aire sous la courbe ROC maximale.

Cet outil de mesure permet d'évaluer nos méthodes de reconnaissance. Pour cela une courbe ROC est construite pour chacune des tâches chirurgicales que l'on cherche à reconnaître. Les vrais positifs représentent les cas où la tâche chirurgicale est bien reconnue. Les faux positifs représentent, quant à eux, les cas où la séquence est reconnue comme appartenant à la tâche chirurgicale ciblée alors qu'elle représente en réalité une autre tâche chirurgicale.

III.4.3 Résultats

La mesure des aires A_z sous la courbe ROC a permis d'évaluer les méthodes mises en place, de comparer les différents choix possibles et de mesurer l'influence de certains paramètres.

III.4.3.1 Mesure de similitude de Piciarelli et al.

Dans un premier temps, l'adaptation de la mesure de similitude de Piciarelli et al. a été évaluée avec différents choix de mesures de distances locales. Les quatre distances locales présentées dans le paragraphe III.3.1 (page 82) : la distance euclidienne, la corrélation, la mesure Chi-Square et la distance de Bhattacharyya**,** ont donc été évaluées. Le paramètre δ, qui contrôle la vitesse de grandissement de la fenêtre glissante a été fixé dans un premier temps avec la valeur proposée par les auteurs de l'article, c'est-à-dire $\delta = 0{,}5$. Les performances sont évaluées avec les histogrammes de mouvements comme signatures visuelles. Les résultats sont présentés dans le Tableau 5.

Les meilleurs résultats ont été obtenus avec la distance Bhattacharyya et permettent même de surpasser la distance DTW. Cela est surprenant car la mesure de similitude de Piciarelli et al. est censée en être une approximation. Il est possible qu'en imposant des bornes sur la fenêtre de recherche, on évite qu'une étape très courte telle que l'incision, par exemple, soit associée à une étape très longue telle que la phacoémulsification.

Tableau 5. Présentation des résultats obtenus avec la distance de Piciarelli et al. pour les quatre types de distances locales ; comparaison avec l'algorithme DTW associé avec la distance Bhattacharyya

	Piciarelli et al.	DTW

	Eucli-dienne	Corréla-tion	Chi-Square	Bhattacharyya	Bhatta-charyya
Incision	0,602	0,500	**0,712**	0,685	0,500
Rhexis	0,500	0,547	0,529	**0,777**	0,500
Hydrodissection	0,523	0,508	**0,746**	0,538	0,524
Phacoémulsifica-tion	0,632	0,698	0,500	0,530	0,679
Epinoyau	0,626	0,500	**0,815**	0,746	0,752
Visqueux	0,500	0,710	0,560	**0,735**	0,645
Mise en Place	0,565	0,623	0,678	**0,745**	0,526
Retrait Visqueux	0,618	0,500	0,505	**0,708**	0,615
Fermeture	0,500	0,695	**0,771**	0,668	0,669
Az moyenne	0,563	0,587	0,646	**0,681**	0,601

Les résultats obtenus avec la distance Bhattacharyya, permettent d'atteindre une aire moyenne sous la courbe ROC de $0{,}681$. Cette valeur est relativement faible. Cela est dû au choix de la valeur $\delta = 0{,}5$, qui n'est pas adaptée à la durée des séquences vidéo de la base et aux variations d'exécution des gestes chirurgicaux. Différentes valeurs de δ ont alors été évaluées. Pour la suite des résultats, la distance locale choisie est la distance de Battacharrya et les résultats obtenus pour les différentes valeurs de δ sont présentés dans le Tableau 6. De même que précédemment, les performances sont évaluées avec les histogrammes de mouvements comme signatures visuelles.

Tableau 6. résultats obtenues avec la distance de Piciarelli et al., avec la distance locale de Bhattacharyya, pour différents choix de facteur de grandissement de la fenêtre glissante

	δ = 0,1	δ = 0,2	δ = 0,3	δ = 0,4	δ = 0,5	δ = 0,8
Incision	0,579	0,603	0,600	0,685	0,685	**0,692**
Rhexis	**0,857**	0,820	0,778	0,617	0,777	0,675
Hydrodissection	0,542	0,561	**0,558**	0,518	0,538	0,525
Phacoémulsification	**0,770**	0,735	0,601	0,598	0,530	0,544
Epinoyau	**0,842**	0,770	0,732	0,725	0,746	0,765
Visqueux	**0,830**	0,811	0,789	0,821	0,735	0,578
Mise en Place	0,697	0,620	0,623	0,598	**0,745**	0,528
Retrait Visqueux	**0,889**	0,817	0,821	0,617	0,708	0,505
Fermeture	**0,743**	0,721	0,715	0,678	0,668	0,638
Az moyenne	**0,750**	0,717	0,691	0,651	0,681	0,605

Les meilleurs résultats ont été obtenus avec une valeur de δ plus faible que celle utilisée par les auteurs. En choisissant une valeur de $\delta = 0{,}1$, une aire moyenne sous la courbe ROC de $0{,}750$

a été mesurée. Cette valeur de δ a donc été utilisée pour la suite des évaluations. Et notamment pour étudier les différents aspects de la caractérisation des vidéos.

III.4.3.2 Influence du choix de la caractérisation des vidéos

Nous avons évalué la mesure de similitude de Piciarelli et al. [44] avec deux types de signatures visuelles. La première est la construction d'histogrammes de mots visuels à partir des descripteurs STIP (paragraphe III.2.1.1.1) et la seconde la construction d'histogrammes de mouvement à partir du flux optique calculé entre deux images consécutives (paragraphe III.2.1.1.2).

Pour les deux types de signatures visuelles, les performances ont été évaluées sur la base de test présentée dans le paragraphe III.4.1. L'aire sous la courbe ROC a été calculée pour chaque tâche chirurgicale. Dans un premier temps, les aires sous la courbe ROC ont été estimées sans effectuer de normalisation sur les vidéos. Puis les effets de chaque type de normalisation ont été évalués indépendamment. Enfin, les résultats ont été calculés avec la combinaison des trois normalisations. Pour la mise à l'échelle des vidéos, le rayon moyen a été fixé empiriquement à 93 pixels, ce qui correspond au rayon moyen mesuré dans la base de données. Le masque circulaire, quant a lui a un rayon de 143 pixels, soit 50 pixels de plus que le rayon de l'iris. Les résultats obtenus en terme d'aire moyenne sous la courbe ROC avec les histogrammes de mouvement pour obtenus pour chacune des tâches chirurgicales sont présenté dans le Tableau 7.

Tableau 7. Résultats en termes d'aire A_z sous la courbe ROC obtenus avec les signatures visuelles en histogrammes de mouvement (HM) construites à partir du flux optique

	HM	HM / REC	HM / ROI	HM / Echelle	HM / Rec Ech ROI
Incision	0,623	**0,700**	0,599	0,637	0,615
Rhexis	0,729	**0,758**	0,651	0,671	0,738
Hydrodissection	0,500	0,640	0,588	**0,642**	0,619
Phacoémulsification	0,742	0,724	0,677	**0,831**	0,821
Epinoyau	0,692	0,764	0,841	0,763	**0,856**
Visqueux	0,956	0,981	0,975	0,966	**0,998**
Mise en Place	0,717	0,631	0,744	0,658	**0,769**
Retrait Visqueux	0,762	0,855	0,786	0,832	**0,860**
Fermeture	0,832	0,768	0,841	0,804	**0,866**
Moyenne	0,728	0,758	0,745	0,756	**0,794**
Erreurs standards	0,042	0,036	0,043	0,038	0,041

Les courbes ROC correspondantes, obtenues pour chacune des tâches chirurgicales, sont présentées dans les Figure 34 et Figure 35. La Figure 34 présente les courbes ROC obtenues sans la normalisation des vidéos et la Figure 35 les courbes ROC obtenues avec la combinaison des 3 types de normalisation.

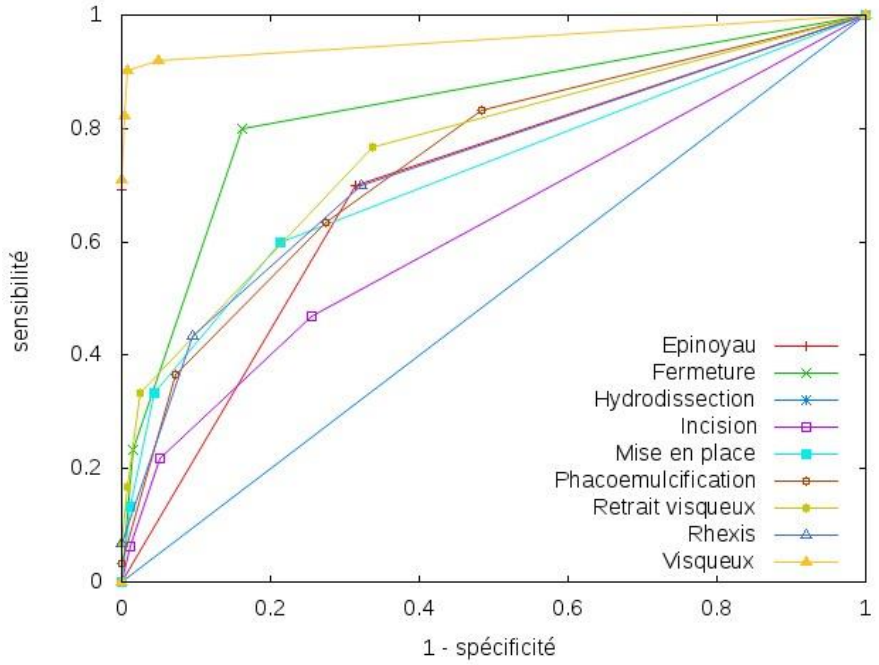


Figure 34. Courbes ROC obtenues sans la normalisation des vidéos pour chaque tâche chirurgicale avec les signatures en histogrammes de mouvement (HM)

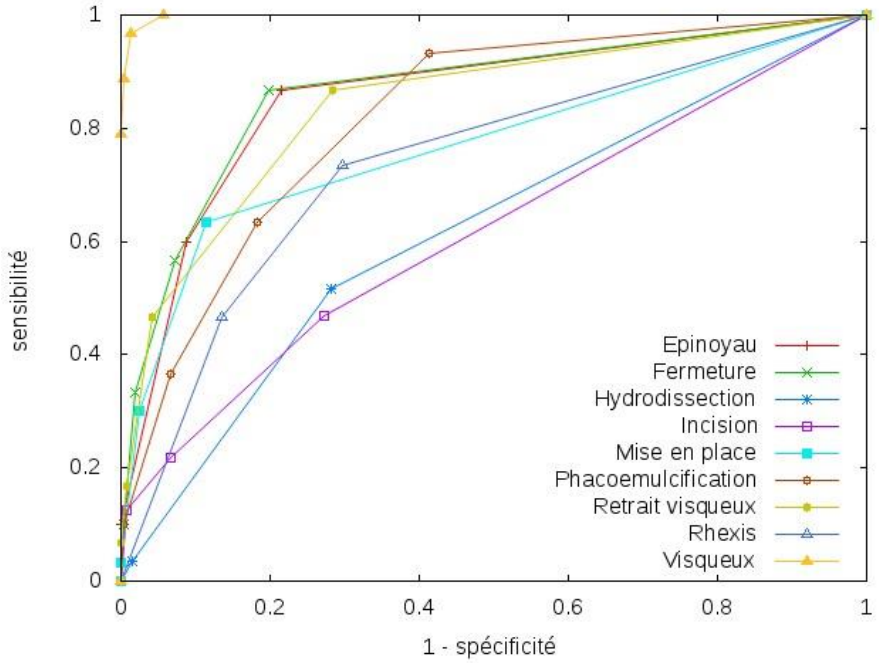


Figure 35. Courbes ROC obtenues avec la normalisation (Recalage + Sélection d'une ROI + Mise à l'échelle) des vidéos pour chaque tâche chirurgicale avec les signatures en histogrammes de mouvement (HM)

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

Chaque type de normalisation permet d'améliorer les performances de la reconnaissance automatique. La reconnaissance des tâches « Incision » et « Rhexis » est essentiellement améliorée par le recalage du centre de l'iris sur le centre de l'image. Nous pouvons donc supposer que les mouvements sont majoritairement induits par le patient et non par les gestes chirurgicaux. La reconnaissance de la tâche « mise en place » en revanche n'est pas améliorée avec le recalage. Dans ce cas, le mouvement de l'œil induit par le geste chirurgical fournit une information pertinente. La sélection d'une région d'intérêt et la mise à l'échelle des images améliorent les performances de reconnaissance de façon plus contrastée pour les différentes tâches chirurgicales. Enfin, pour toutes les tâches chirurgicales les résultats sont améliorés par la combinaison des trois normalisations, à l'exception de la reconnaissance de la tâche « Incision » pour laquelle l'aire sous la courbe ROC est légèrement inférieure avec normalisation.

Dans le cas de l'utilisation des histogrammes de mots visuels comme signatures des séquences vidéo, une étape d'apprentissage est nécessaire pour la construction du dictionnaire de mots visuels. Pour cela la base d'apprentissage de 30 vidéos a été utilisée. Un dictionnaire de 1000 mots visuels a été construit et la taille du rayon moyen et du masque de la ROI sont respectivement de 93 et 143 pixels, comme pour l'utilisation des histogrammes de mouvement comme signatures visuelles. Les résultats en termes d'aire sous la courbe ROC obtenus avec la base de test pour chacune des tâches chirurgicales sont présentés dans le Tableau 8.

Tableau 8. Résultats en termes d'aire A_z sous la courbe ROC obtenus avec les signatures en histogrammes de mots visuels (BoW) construit à partir des descripteurs STIP

	BoW	BoW / Rec	BoW/ ROI	BoW/ Echelle	BoW/ Rec ROI Echelle
Incision	0,724	0,728	**0,808**	0,757	0,703
Rhexis	**0,946**	0,907	0,913	0,932	0,917
Hydrodissection	0,621	0,605	0,652	**0,678**	0,624
Phacoémulsification	**0,882**	0,858	0,859	0,858	0,828
Epinoyau	**0,900**	0,817	0,867	0,783	0,650
Visqueux	0,998	0,998	0,998	0,998	0,998
Mise en Place	0,722	**0,784**	0,779	0,742	0,775
Retrait Visqueux	**0,842**	0,754	0,814	0,816	0,802
Fermeture	0,798	0,826	0,845	**0,848**	0,845
Moyenne	0,826	0,809	**0,837**	0,824	0,793
Erreurs standards	0,040	0,037	0,032	0,033	0,041

Les courbes ROC correspondantes, obtenues pour chacune des tâches chirurgicales, sont présentées dans les Figure 36 et Figure 37. La Figure 36 présente les courbes ROC obtenues sans la normalisation des vidéos et la Figure 37 les courbes ROC obtenues avec la normalisation qui donne le meilleur résultat en terme d'aire moyenne sous la courbe ROC : la sélection d'une région d'intérêt.

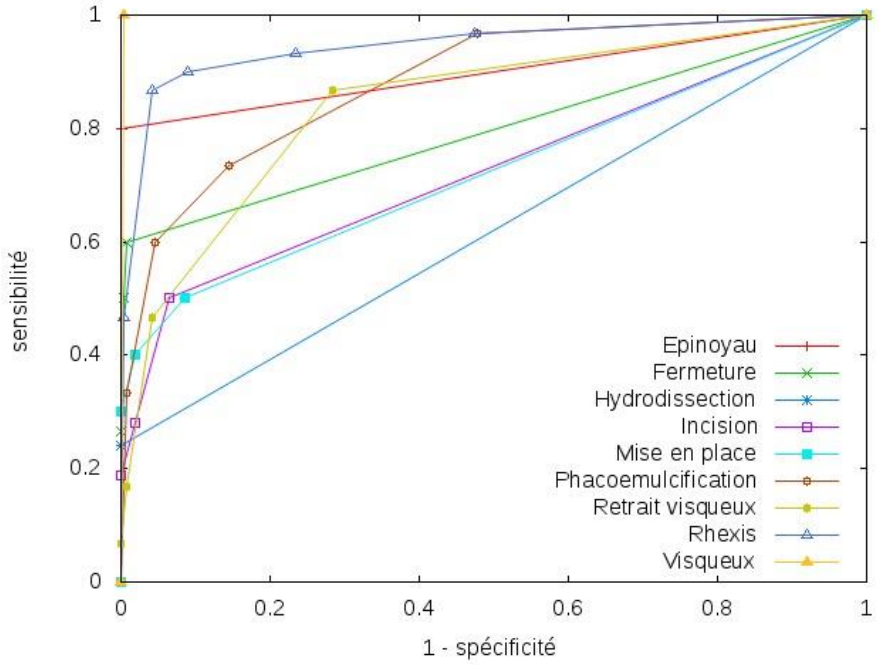


Figure 36. Courbes ROC obtenues sans normalisation des vidéos pour chaque tâche chirurgicale avec les signatures en histogrammes de mots visuels (BoW) construit à partir des descripteurs STIP

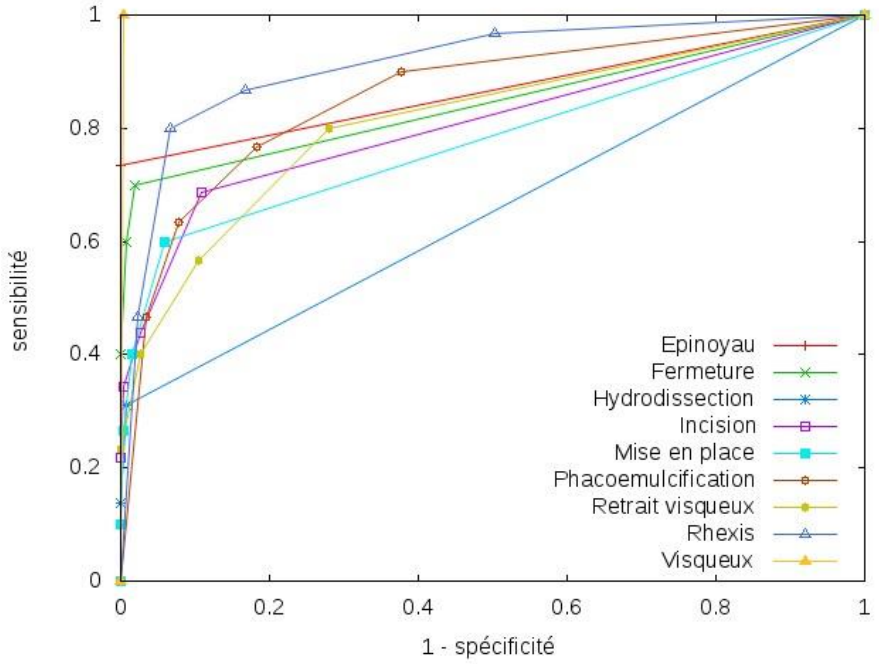


Figure 37. Courbes ROC obtenues avec la normalisation des vidéos (sélection d'une ROI) pour chaque tâche chirurgicale avec les signatures en histogrammes de mots visuels (BoW) construit à partir des descripteurs STIP

Les performances obtenues avec les signatures sous forme d'histogrammes de mots visuels sont supérieures à celles obtenues avec les histogrammes de mouvement, avec une aire moyenne sous la courbe ROC de $0{,}826$ sans normalisation des vidéos. La sélection d'une région d'intérêt a

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

permis d'améliorer les performances de reconnaissance (A_z moyenne $= 0{,}837$). En revanche, la mise à l'échelle des vidéos n'a pas eu d'effets sur les performances de reconnaissance. Cela peut s'expliquer par le fait que les descripteurs spatio-temporels sont invariants aux changements d'échelle. Le recalage des images n'améliore pas les performances de reconnaissance bien qu'il supprime les mouvements parasites qui affectent l'extraction des STIP. Cela peut s'expliquer par le fait que la représentation par sacs de mots visuels ne prend pas en considération la position des mouvements détectés, contrairement à la représentation en histogrammes de mouvements.

III.4.3.3 Temps de calcul

L'objectif étant de mettre en place des méthodes rapides, compatibles avec le temps réel, les temps de calcul ont été mesurés pour évaluer l'impact du choix des signatures visuelles et des étapes de normalisation. Les résultats sont présentés dans le Tableau 9.

Tableau 9. Temps de calcul en secondes pour les différents éléments de la caractérisation des vidéos

	Sans normalisation	Recalage	Echelle	ROI	Rec-ROI-Echelle
Normalisation	0	0.0206	0.054	0.02	0.0299
Calcul flux optique	0.0176	0.0031	0.0031	0.0032	0.001
Extraction STIP	1.019	0.3173	0.3145	0.3347	0.1561
Calcul BoW	0.0049	0.0029	0.003	0.0033	0.0017

Les temps de calcul moyens, obtenus avec un processeur « quad core » Intel(R) Core(TM) i7-3770 (3.40GHz), pour les étapes de normalisation se situe entre 20 ms et 54 ms. Cependant, le fait de normaliser les vidéos engendre une accélération des calculs pour l'extraction des caractéristiques. En effet chaque type de normalisation permet de diminuer d'un tiers les temps de calcul pour l'extraction des descripteurs STIP et le calcul du flux optique. On peut également constater que, si l'utilisation des histogrammes de mots visuels pour caractériser les séquences vidéo permet d'obtenir des performances de classification supérieures à l'utilisation des histogrammes de mouvement, l'extraction des descripteurs STIP est plus coûteuse en temps de calcul et peu compatible avec une utilisation en temps réel. Cela s'explique par le fait que la recherche des points d'intérêts ne se fait pas uniquement dans le domaine spatial comme dans le cas du détecteur de singularités utilisé pour l'extraction du flux optique, mais également dans le domaine temporel, par extension du détecteur de Harris avec une composante temporelle.

III.5 Discussion – Conclusion

Les résultats obtenus sont satisfaisants au regard de la rapidité de l'algorithme. Les meilleurs résultats ont été obtenus avec la méthode de caractérisation des vidéos de l'état de l'art, c'est-à-dire l'utilisation des histogrammes de mots visuels. Dans ce cas, l'aire moyenne sous la courbe ROC est de $0{,}826$ sans normalisation des vidéos, et de $0{,}837$ avec la restriction de l'extraction des caractéristiques à une région d'intérêt. Mais, on a vu dans le Tableau 9 que l'extraction des descripteurs STIP, permettant de construire ces signatures visuelles, n'était pas compatible avec une utilisation en temps réel. Cela s'explique par l'extension dans le domaine temporel de la recherche de singularités par le détecteur de Harris, contrairement au détecteur de singularités utilisé pour l'extraction du flux optique. La méthode que nous choisissons, bien que moins performante, fournit néanmoins des résultats satisfaisants tout en étant compatible avec les contraintes de temps réel imposées par nos objectifs. L'aire moyenne sous la courbe ROC obtenue sans étape de normalisation est de $0{,}728$. Ce score est amélioré par l'utilisation des étapes de normalisation, et l'aire moyenne sous la courbe ROC atteint une valeur de $0{,}794$ lorsqu'une combinaison des trois normalisations est utilisée.

Les étapes de normalisation ont ainsi prouvé leur efficacité, en améliorant les performances de reconnaissance. Bien que nécessitant de 20 à 54 ms de temps de calcul, l'utilisation des étapes de normalisation permet de réduire les temps de calcul de l'extraction des caractéristiques visuelles. En revanche, l'impact du choix du rayon sur les performances de reconnaissance n'a pas été étudié. Il serait intéressant d'étudier ce point, car dans le cas de la tâche « Incision », par exemple, la position des instruments et des gestes chirurgicaux est majoritairement située en bordure de l'iris et une ROI trop petite peut engendrer une perte d'information.

L'adaptation de la mesure de similitude proposée par Piciarelli et al. [44], pour laquelle la mesure de distance locale entre deux points, initialement une distance euclidienne, a été remplacée par la distance de Bhattacharyya, a également prouvé son efficacité pour la comparaison de vidéos chirurgicales. Les bons résultats obtenus avec la mesure de distance de Bhattacharyya s'expliquent par le fait que chaque image de la vidéo est caractérisée par un ou plusieurs histogrammes de répartition. Or la distance de Bhattacharyya est réputée meilleure que la distance euclidienne pour la mise en correspondance d'histogrammes.

Cette approche a été évaluée en s'appuyant sur une base de séquences ou chaque séquence représente une tâche chirurgicale. Le problème initial, de reconnaître en direct, à tout moment de la chirurgie, quelle tâche chirurgicale est réalisée par le chirurgien, a donc été simplifié pour évaluer les méthodes de reconnaissance. Nous avons cherché ici à reconnaître la tâche chirurgicale représentée dans la séquence requête. Pour la suite de ce travail de thèse, nous avons cherché à segmenter en direct et en temps réel une vidéo de chirurgie de la cataracte complète. Pour cela nous avons cherché à nous appuyer sur une description précise et complète de la chirurgie et une modélisation statistique du processus chirurgical.

Pour conclure, Le système de reconnaissance automatique par recherche de cas similaires proposé a fait ces preuves et valide les contraintes de temps réel qui nous sont imposées. Il est maintenant nécessaire d'être capable de séquencer temporellement une vidéo de chirurgie en gestes chirurgicaux.

Dans le chapitre suivant nous présentons et évaluons deux méthodes de séquençage automatique d'une vidéo d'opération de la cataracte, du début à la fin de la chirurgie (vidéo complète). Ces méthodes s'appuient sur des modèles statistiques du déroulement de la chirurgie.

Chapitre IV. Séquençage multi-échelles d'une vidéo de chirurgie

IV.1 Modélisation du processus chirurgical 98
IV.1.1 Modèles Graphiques 99
IV.1.1.1 Les graphes quelconques 99
IV.1.1.2 Les arbres 100
IV.1.2 Information contextuelle en analyse de vidéos médicales 101
IV.1.2.1 Construction d'une chirurgie moyenne 101
IV.1.2.2 Modèles Statistiques utilisés en analyse de vidéos médicales 102
IV.1.2.2.1 Modèles Markoviens 102
IV.1.2.2.2 Les Champs Markoviens conditionnels 104
IV.1.2.2.3 Les Systèmes Linéaires Dynamiques 105
IV.1.2.3 Modélisation des relations entre les niveaux 105
IV.1.2.3.1 Arbres de décision 106
IV.1.2.3.2 Réseaux bayésiens 106
IV.1.3 Synthèse 107

IV.2 Construction d'arbres (Piciarelli et al.) 109
IV.2.1.1 Méthode de Piciarelli et al. 109
IV.2.2 Construction de l'arbre 111
IV.2.2.1 Méthode non supervisée 111
IV.2.2.2 Méthode supervisée 112
IV.2.3 Inférence de l'arbre 113
IV.2.4 Résultats 114
IV.2.4.1 Méthode non supervisée 114
IV.2.4.2 Méthode supervisée 115
IV.2.5 Synthèse 117

IV.3 Modélisation statistique multi-échelles 119
IV.3.1 Construction du modèle 119
IV.3.1.1 Réseau Bayésien 120
IV.3.1.2 HMM 123
IV.3.1.1 CRF 123
IV.3.1.2 Retour du HMM vers le réseau bayésien 125
IV.3.2 Caractérisation de la vidéo 126
IV.3.2.1 Structure de l'analyse 126
IV.3.2.2 Génération des observations 126
IV.3.2.2.1 Présence des instruments 127
IV.3.2.2.2 Analyse du mouvement 127
IV.3.3 Evaluation 128
IV.3.3.1 Optimisation des paramètres 129
IV.3.3.1.1 Utilisation des HMM 129

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

IV.3.3.1.2 Utilisation des CRF 131
IV.3.3.2 Inférence 132
IV.3.3.2.1 Réseau bayésien seul 132
IV.3.3.2.2 HMM 133
IV.3.3.2.3 CRF 133
IV.3.3.3 Résultats 134
IV.3.3.3.1 Réseau Bayésien 134
IV.3.3.3.2 Réseau Bayésien et HMM 135
IV.3.3.3.3 Réseau Bayésien et retour HMM « phases » 138
IV.3.3.3.4 Réseau Bayésien et CRF 139
IV.3.3.4 Conclusion 140

IV.4 Discussion – Conclusion 142

Nous avons montré dans le Chapitre III que nous étions capables de reconnaître avec de bonnes performances la tâche chirurgicale représentée dans une séquence vidéo requête (paragraphe III.4.3). Nous allons maintenant chercher à reconnaître à chaque instant de la chirurgie quelle tâche chirurgicale est réalisée. Pour cela nous ne disposons plus de séquences segmentées pour lesquelles une séquence représente une tâche chirurgicale. Nous pourrions alors, par exemple, considérer la vidéo comme une succession de sous-séquences de tailles fixes et appliquer telles quelles les méthodes de reconnaissance automatiques de tâches chirurgicales développées précédemment. Cependant, comme l'a démontré Lalys [23] dans sa thèse, l'approche consistant à classifier indépendamment chacune des images ou séquences qui composent la vidéo ne permet pas d'obtenir les résultats escomptés. L'introduction d'information contextuelle, apportée par la connaissance du processus chirurgical, permet d'améliorer les performances de reconnaissance, et de segmentation automatique d'une vidéo de chirurgie.

Le principe général de la méthode de séquençage automatique d'une vidéo de chirurgie requête est présentée dans la Figure 38.

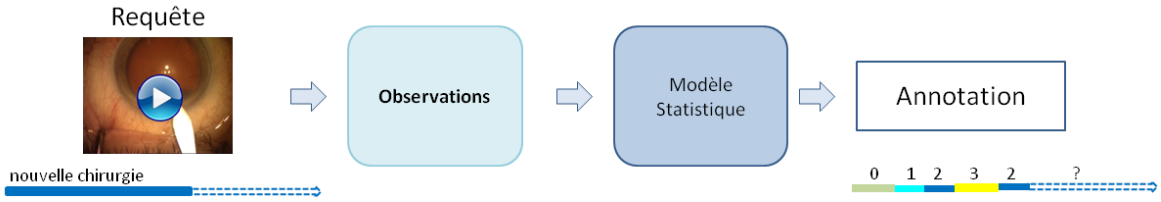


Figure 38. Principe général de notre méthode de séquençage automatique d'une chirurgie requête

Nous étudierons dans un premier temps les méthodes de la littérature relatives à la modélisation du processus chirurgical. Puis nous présenterons deux approches que nous avons mises en place : la construction d'une chirurgie moyenne via un arbre et la construction d'un modèle statistique multi-échelle.

IV.1 Modélisation du processus chirurgical

Il est possible, grâce à des méthodes de fouille de données d'extraire de la connaissance à partir d'une base de cas archivés. Le raisonnement à base de cas et la recherche de vidéos par le contenu (CBVR) offrent une première réponse à notre problème de reconnaissance automatique des tâches chirurgicales par une recherche des plus proches voisins dans la base de données. L'utilisation d'autres méthodes de fouille de données permet d'aller plus loin en construisant par apprentissage un modèle statistique du processus chirurgical. Ce dernier permet d'utiliser les connaissances apprises de la base de données pour apporter une information contextuelle lors de la reconnaissance automatique des tâches chirurgicales, en modélisant les relations qui existent entre les différentes tâches. Il existe dans la littérature différents types de modélisation du processus chirurgical, dont une grande partie s'appuie sur des modèles graphiques.

IV.1.1 Modèles Graphiques

On cherche à modéliser les relations qui existent entre les différentes tâches chirurgicales qui décrivent la chirurgie de la cataracte, ainsi que les relations qui les lient aux observations (signatures visuelles). Pour cela, la construction de modèles graphiques est une approche intuitive et visuelle. De manière générale, un graphe permet de représenter les relations entre les différents éléments qui composent un système. Il existe différents types de modèles graphiques, que l'on peut regrouper en deux grandes classes : les arbres et les graphes quelconques. Les principales notions relatives aux modèles graphiques sont présentées ici.

IV.1.1.1 Les graphes quelconques

Dans le cadre des modèles graphiques, les différents éléments du système (dans notre cas les tâches chirurgicales par exemple) sont représentés par des **nœuds** (ou **sommets**). Les relations entre ces éléments sont représentées par des **arcs** (ou **arêtes**). Un graphe G est alors un couple $(\mathcal{V}, \mathcal{E})$, où $\mathcal{V} = \{S_1, \dots, S_n\}$ est l'ensemble des nœuds du graphe et $\mathcal{E} = \{e_1, \dots, e_m\}$ l'ensemble des arcs (sous-ensemble de $\mathcal{V} \times \mathcal{V}$). Un graphe est d'ordre n s'il comporte n sommets et un arc e du graphe est une paire de nœuds $e = (S_i, S_j)$, où S_i et S_j sont les extrémités de l'arc. Les sommets S_i et S_j sont alors dits adjacents dans G. Le parcours d'un graphe se fait en se déplaçant de nœud en nœud, via les arcs. L'ensemble des nœuds et arcs ainsi parcourus forment une chaîne. Une chaîne est une liste $w = \{S_1, \dots, S_k\}$ tel qu'il existe un arc entre chaque paire de nœuds (S_i, S_{i+1}) successifs. Si chacun des arcs n'est parcouru qu'une seule fois, on parle de chaîne simple. Un cycle est une chaîne simple rebouclant sur elle-même. On parle alors de graphe **acyclique** lorsque celui-ci ne comporte aucun cycle.

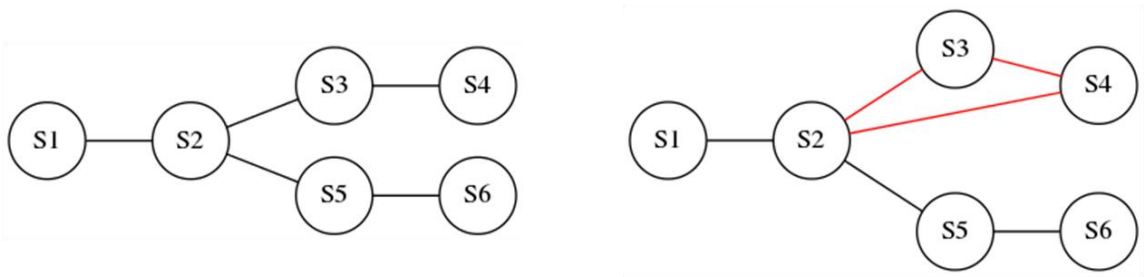


Figure 39. A gauche un exemple de graphe non orienté acyclique; A droite, un exemple de cycle

On parle de graphe **non-orienté**, lorsque l'on ne distingue pas l'extrémité initiale de l'extrémité finale. Dans le cas d'un graphe **orienté**, un arc e du graphe est une paire de sommets $e = (S_i, S_j)$, où S_i est l'extrémité initiale et S_j l'extrémité finale. Deux exemples de graphes orientés et non orientés sont présentés dans la Figure 40.

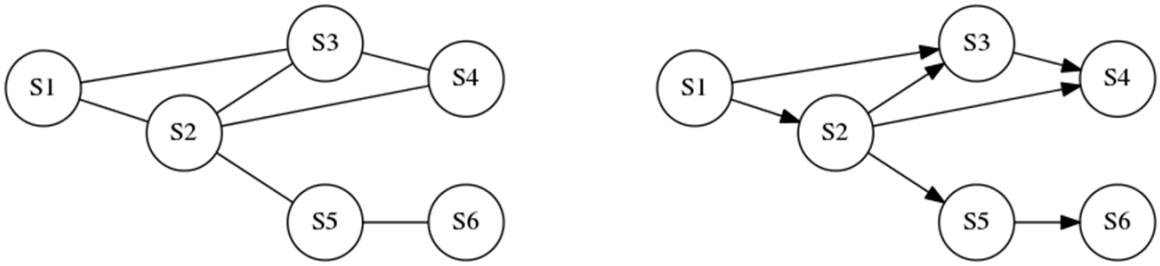


Figure 40. A gauche un exemple de graphe non orienté; A droite, un exemple de graphe orienté

Dans le cas des graphes orientés, on parle alors de chemin et non pas de chaîne et l'on ne peut pas prendre les arcs à rebours. Si un graphe est non orienté et acyclique il s'agit d'une forêt, c'est-à-dire un ensemble d'arbres.

IV.1.1.2 Les arbres

Un **arbre** est un graphe connexe sans cycle. On parle de graphes connexes si depuis un sommet il existe une chaîne pour atteindre tout autre sommet. Les graphes non connexes sont composés d'un ensemble de graphes, appelés « composantes connexes ». Dans le cas des graphes sans cycle, chacune de ses composantes connexes est acyclique. Il s'agit donc bien d'un ensemble d'arbres que l'on nomme alors forêt.[2]

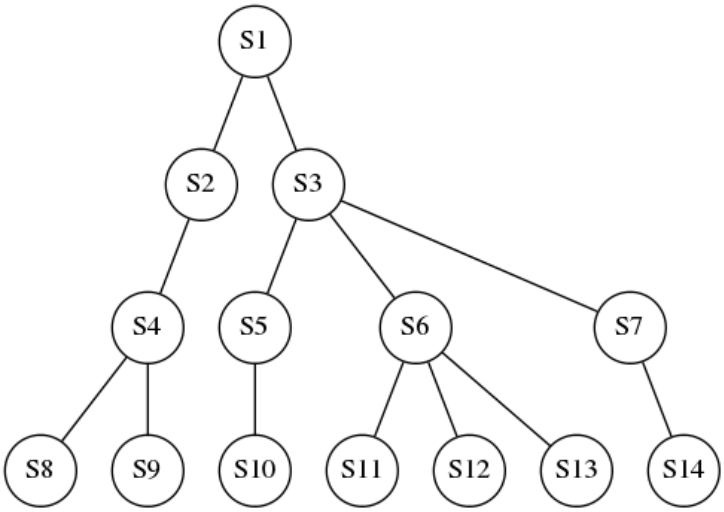


Figure 41. Exemple d'arbre

Dans le cas des arbres la notion de **racine** est introduite. Pour les graphes orientés, la racine est le sommet qui ne possède pas de prédécesseur dans l'arbre. Dans le cas des graphes non orientés, le choix est arbitraire. On parle alors ensuite de nœuds **parents** et de nœuds **fils**. Le nœud parent d'un nœud S est le prédécesseur (unique) de S sur le chemin de la racine à S. La racine, elle, est le seul nœud de l'arbre qui n'a pas de parents. Les nœuds fils sont les nœuds adjacents à S, autres que son nœud parent. Enfin, un nœud qui n'a pas de fils est une **feuille**. Dans l'exemple

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

d'arbre de racine S_1 proposé dans la Figure 41, les nœuds S_8 à S_{14} sont des feuilles de l'arbre. Le parent du nœud S_6 est le nœud S_3, et les nœuds S_5, S_6 et S_7 sont les nœuds fils du nœud S_3.

Les arbres et les graphes s'avèrent être de bons outils pour manipuler des objets et leurs relations. C'est pourquoi la majorité des méthodes, présentées ci-après (paragraphe IV.1.1), qui permettent d'apporter une information contextuelle, s'appuie sur ces outils.

IV.1.2 Information contextuelle en analyse de vidéos médicales

IV.1.2.1 Construction d'une chirurgie moyenne

Une première idée consiste à s'appuyer sur l'algorithme DTW présenté dans le paragraphe III.1.2.1. Cet algorithme a été développé pour mesurer la similitude entre deux séries temporelles. Pour cela, l'algorithme recale temporellement les deux séries, en compensant les distorsions existantes. Ce recalage est utilisé par Blum et al. [24], Padoy et al. [25] ou Lalys et al. [22] pour construire une chirurgie moyenne à partir des cas de la base d'apprentissage en s'appuyant sur la méthode de Wang et Gasser [77]. La chirurgie requête est ensuite recalée sur la chirurgie moyenne via sa signature visuelle. Une fois recalées, les phases de la chirurgie moyenne sont transposées sur la chirurgie requête (Figure 42).

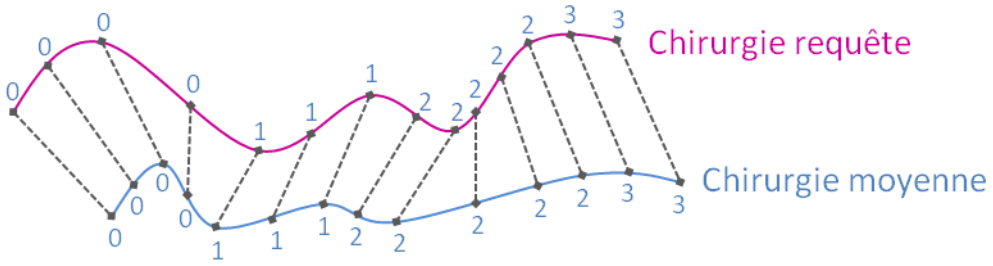


Figure 42. Exemple de segmentation d'une chirurgie requête en la recalant sur une chirurgie moyenne via l'algorithme DTW

Cette méthode donne de bons résultats et surpasse même la méthode de modélisation de la chirurgie par une chaîne de Markov cachée [22] [24] [25]. Cependant son premier inconvénient est qu'elle nécessite d'enregistrer l'intégralité de la chirurgie pour pouvoir la recaler sur la chirurgie moyenne. Cette méthode n'est donc pas compatible avec une analyse en direct de la vidéo. De plus, cette méthode ne supporte qu'un ordonnancement unique des phases chirurgicales. En effet, pour construire la chirurgie moyenne, l'enchaînement des phases chirurgicales doit être identique quelle que soit la chirurgie. Cette méthode n'est donc pas compatible avec la description en tâches chirurgicales avec laquelle nous travaillons, présentée dans le paragraphe II.2.2.1. En effet cette description permet de multiples ordonnancements des tâches chirurgicales. Il est donc nécessaire de choisir un modèle qui prenne en compte cet aspect.

Il existe différents modèles statistiques permettant de gérer les différents ordonnancements. La construction et l'utilisation d'un modèle statistique du processus chirurgical permet une meilleure analyse des vidéos de chirurgies, en apportant une information contextuelle. En effet, il est peu probable qu'une chirurgie commence par une tâche de suture, ou qu'une incision est lieu en fin de chirurgie. Il existe un déroulement probable de la chirurgie, mais qui peut comporter des

variations selon les chirurgies. C'est cela que l'on va chercher à modéliser, en quantifiant les relations qui existent entre les différentes tâches chirurgicales, tel que les probabilités de transition d'une tâche vers une autre par exemple. Il existe dans la littérature, différentes solutions proposées pour modéliser le processus chirurgical.

IV.1.2.2 Modèles Statistiques utilisés en analyse de vidéos médicales

Les modèles statistiques utilisés en analyse de vidéos médicales sont issus des méthodes d'analyses de séries temporelles et permettent de modéliser le déroulement temporel de la chirurgie.

IV.1.2.2.1 Modèles Markoviens

Les **chaînes de Markov** permettent d'étudier un système discret à N états différents $\{S_1, S_2, \dots, S_n\}$ qui évolue au cours du temps. A chaque pas de temps, le système décide s'il change d'état ou s'il reste dans son état actuel. Le système possède la propriété de Markov, c'est-à-dire que la prédiction de l'état q au temps t ne dépend que de l'état au temps $(t-1)$. C'est-à-dire :

$$P(q_t = S_j | q_{t-1} = S_i)$$

Les informations supplémentaires concernant le passé ne permettent pas de rendre la prédiction plus précise. Les probabilités de transition d'états ne changent pas au cours du temps et sont définies par apprentissage. Une matrice de transition $\boldsymbol{A} = (a_{ij})_{0 \le i,j < N}$ est construite, et définie de la manière suivante :

$$a_{ij} = P(q_t = S_j | q_{t-1} = S_i)$$

avec $a_{ij} \geq 0$ et $\sum_{j=1}^{N} a_{ij} = 1$. Chaque élément de la matrice a_{ij} de la matrice représente la probabilité de transition vers l'état S_j, sachant que l'on est dans l'état S_i.

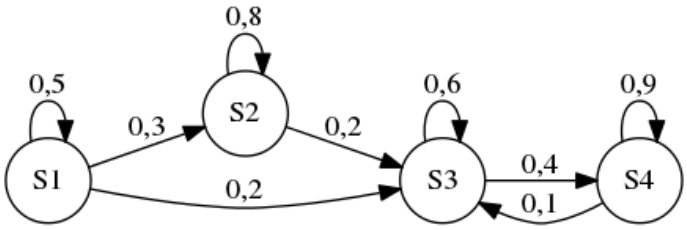


Figure 43. Exemple de diagramme de changement d'état à 4 états

Un exemple de diagramme de changement d'état est présenté dans la Figure 43. Dans cet exemple, la probabilité d'obtenir la séquence $\mathcal{Y} = \{S_1, S_1, S_2, S_3, S_4, \}$, par exemple, est calculée de la manière suivante :

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

$$P(\mathcal{Y}|A) = P(S_1).P(S_1|S_1).P(S_2|S_1).P(S_3|S_2).P(S_4|S_3)$$

Les chaînes de Markov sont particulièrement adaptées à notre problème. Chaque tâche chirurgicale correspond à un état du système. Cependant, dans notre problème, les états ne sont pas directement observables. Les modèles de Markov cachés semblent alors plus adaptés.

Les **modèles de Markov cachés** (MMC, ou HMM pour « Hidden Markov Models » en anglais) sont très fréquemment utilisés dans la littérature [27][34][41]. Ils modélisent un processus markovien, pour lequel les états ne sont pas observables (ils sont cachés), seuls les événements qu'ils produisent sont observables. Ces observations génèrent des vecteurs de probabilités qui permettent d'étudier les états cachés. Le modèle se compose alors de deux processus : un processus à N états non observables et un processus observable. Un modèle de Markov caché est donc défini par un quadruplet $\{\mathcal{S}, \boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B}\}$:

- $\mathcal{S} = \{S_1, S_2, \dots, S_n\}$ représente les états cachés du système
- $\boldsymbol{\pi} = \{p_1, p_2, \dots p_n\}$ est le vecteur des probabilités initiales où p_i est la probabilité que l'état S_i soit l'état initial
- $\boldsymbol{A} = \left(a_{ij}\right)_{0 \le i,j < N}$ est la matrice des transitions et a_{ij} représente la probabilité de transition de l'état S_i vers l'état S_j
- $\boldsymbol{B}$ est la matrice d'observation, ou $b_i(k) = P(o_t = \mathrm{k} \,|q_t = S_i)$ la probabilité d'émettre le symbole k étant dans l'état S_i

Un exemple de modèle de Markov à quatre états cachés est présenté dans la Figure 44. Les nœuds $\{S_1, S_2, S_3, S_4\}$ représentent les états cachés et les valeurs j et k sont les valeurs de sorties.

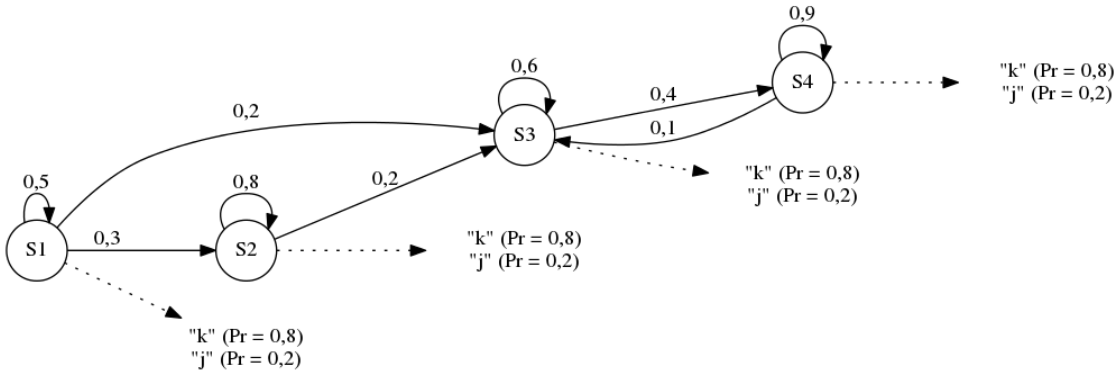


Figure 44. Exemple de modèle de Markov à 4 états cachés ; les flèches en pointillés indiquent les sorties probables à chaque passage dans un état

Dans le cas de l'analyse automatique du processus chirurgical, les tâches chirurgicales ne sont pas directement observées (états cachés). Seules les caractéristiques visuelles, ou la présence des instruments par exemple, sont observables (valeurs de sorties). Les modèles de Markov cachés permettent, connaissant le modèle, de trouver la séquence la plus probable de tâches (états cachés) ayant conduit à la génération d'une séquence de sortie donnée. Cela se résout avec l'algorithme de Viterbi [78]. Une limite des chaînes de Markov cachées est qu'elles nécessitent que la base d'apprentissage représente tous les cas de transition qu'il est possible de rencontrer. Si un

cas n'est pas présent dans la base d'apprentissage, alors le modèle retiendra qu'une telle transition est impossible. Si le cas est présent dans la chirurgie requête, alors la transition ne sera pas reconnue.

IV.1.2.2.2 Les Champs Markoviens conditionnels

Les Champs Markoviens conditionnels ou « Conditional Random Fields » en anglais sont une alternative aux modèles de Markov cachés. Ils ont été introduits par Lafferty et al. pour l'annotation de séries temporelles [79] et ils semblent donner de meilleurs résultats que les HMM pour l'analyse automatique de vidéos chirurgicales [30] et [51].

Le modèle est un graphe probabiliste qui modélise par une distribution log linéaire les séquences de labels en fonction d'une séquence d'observations donnée. Comme le montre la Figure 45, les CRF appartiennent à la famille des modèles graphiques non dirigés.

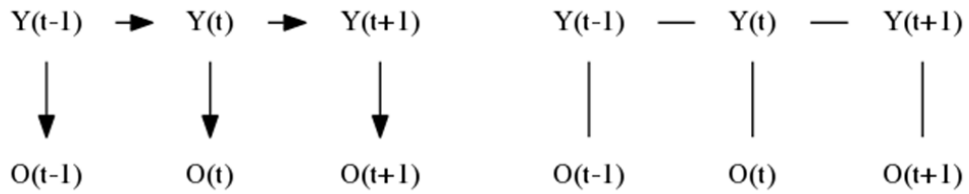


Figure 45. A gauche, exemple de structure d'une chaîne de Markov cachée simple ; à droite, exemple de structure en chaîne d'un CRF

L'objectif est de labelliser une nouvelle observation o_t en choisissant le label y_t qui maximise la probabilité $P(y_t|o_t)$. Soit $\mathcal{S} = \{S_1, S_2, \dots, S_n\}$ l'ensemble des labels possibles et $\mathcal{Y} = \{y_t\}_{0<t\leq T}$ une séquence de labels avec $y_t \in \mathcal{S}$. La distribution des probabilités conditionnelles de la séquence de labels $\mathcal{Y}$ étant donné la séquence d'observations $\mathcal{O} = \{o_t\}_{0<t\leq T}$ peut être modélisée de la façon suivante :

$$P(\mathcal{Y}|\mathcal{O}) \propto \exp\left(\sum_{t=1}^{T} \boldsymbol{\lambda}_t \psi_t^u(y_t, \boldsymbol{o}_t) + \sum_{t=1}^{T-1} \boldsymbol{\mu}_t \psi_{t,t+1}^b(y_t, y_{t+1}, \boldsymbol{o}_t)\right)$$

Les fonctions caractéristiques ψ^u et ψ^b sont appelées des potentiels « unaires » et « binaire ». Elles sont pondérées par les poids $\boldsymbol{\lambda}$ et $\boldsymbol{\mu}$, calculés par apprentissage à partir de l'ensemble des couples $(y_t, \boldsymbol{o}_t)$. Les potentiels « unaires » donnent le score d'assignation d'un label à une observation. Alors que les potentiels « par paire » représentent, quant à eux, la probabilité de passer d'un label x_i à un label x_j quand on passe du pas de temps t au pas de temps $t+1$.

On peut considérer les CRF comme une généralisation des HMM. Un HMM est un cas particulier de CRF où des probabilités constantes sont utilisées pour modéliser des transitions d'état. Le premier avantage des CRF est leur nature conditionnelle, qui les rend moins dépendants des suppositions d'indépendance exigées par les HMM. Ils sont, de plus, moins sensibles au problème du biais du label, exposé par Lafferty et al. [79]. Les CRF permettent de mieux gérer le manque d'information éventuelle de la base de données. Enfin, les CRF permettent de prendre en compte

différentes sources d'observations, ce qui n'est pas le cas avec les HMM. Effectivement, dans le cas des HMM, o_t est un scalaire, alors que dans le cas des CRF, il s'agit d'un vecteur de taille quelconque. En revanche, contrairement aux HMM, l'implémentation des CRF est complexe.

IV.1.2.2.3 Les Systèmes Linéaires Dynamiques

Les systèmes linéaires dynamiques (LDS, pour « Linear Dynamical System» en anglais) ont été utilisés par Haro et al. [29] et Zappella et al.[31]. Ils permettent également de modéliser une série temporelle. Zappella et al. montrent comment il est possible de modéliser les observations émises par chaque geste chirurgical comme la sortie d'un LDS [31]. Comme pour les modèles précédents, nous noterons $\boldsymbol{o}_t$ le signal observé au temps t et $\boldsymbol{y}_t$ est l'état caché. L'observation $\boldsymbol{o}_t$ est considérée comme la sortie d'un LDS :

$$\begin{cases} \boldsymbol{y}_{t+1} = \boldsymbol{A}\boldsymbol{y}_t + \boldsymbol{B}\boldsymbol{u}_t \\ \boldsymbol{o}_t = \boldsymbol{C}\boldsymbol{y}_t + \boldsymbol{w}_t \end{cases}$$

La matrice $\boldsymbol{B}$ est une matrice de bruit coloré qui permet de mettre en relief les corrélations du bruit $\boldsymbol{u}_t \overset{iid}{\sim} \mathcal{N}(0, \boldsymbol{I})$ du processus (un bruit gaussien, de moyenne nulle et de matrice de covariance identité). La matrice $\boldsymbol{A}$ est la matrice de transition d'états et $\boldsymbol{C}$ est la matrice des relations entre les observations et les états cachés. La mesure du bruit d'une séquence $\boldsymbol{w}_t$ est également un bruit gaussien, de moyenne zéro et de matrice de covariance $\boldsymbol{R}$, c'est-à-dire $\boldsymbol{w}_t \overset{iid}{\sim} \mathcal{N}(0, \boldsymbol{R})$. Un modèle LDS est alors défini par le quadruplet $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{R})$. Un modèle va être construit pour chaque geste chirurgical. Cependant cette représentation n'est pas unique et deux modèles $\mathcal{M} = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{R})$ et $\mathcal{M}' = (\boldsymbol{T^{-1}AT}, \boldsymbol{T^{-1}B}, \boldsymbol{CT}, \boldsymbol{R})$ sont identiques par un changement de coordonnées des variables d'état $\boldsymbol{y}_t = \boldsymbol{T}\boldsymbol{y}'_t$ (où $\boldsymbol{T}$ est une matrice inversible). Il est important de prendre cela en considération lors de la comparaison de deux modèles entre eux afin de classifier les gestes chirurgicaux.

IV.1.2.3 Modélisation des relations entre les niveaux

Plus le niveau de granularité est élevé, plus la reconnaissance a de chances d'être aisée. Cela est lié au fait que, dans le cas de phases chirurgicales par exemple, celles-ci sont très différentes les unes des autres, donc facilement différenciables. De plus, elles sont toujours présentes et leur ordonnancement est toujours le même, cela permet d'avoir une modélisation du processus chirurgical simple. Toutes les relations d'indépendances sont bien modélisées par le modèle de Markov et il n'existe pas de relations possibles entre phases autres que celles présentes dans la base de données. Cependant ce niveau de description, même s'il apporte une première information, ne permet pas de décrire la chirurgie de façon suffisamment précise pour permettre de détecter des cas anormaux ou d'apporter une aide ciblée aux chirurgiens. Il est donc nécessaire de travailler à un niveau de granularité plus fin. Cependant dans ce cas, les étapes ou les activités sont moins facilement différentiables. Par exemple les étapes « Cracking » et « Sillons » qui permettent de séparer le cristallin en morceaux lors de la phacoémulsification sont très proches visuellement. De plus les relations entre les étapes, ou les activités, sont multiples, ce qui rend le modèle complexe. Cela est lié au fait que certaines étapes ou activités sont réalisées plusieurs fois tout au long de la

chirurgie, dans des ordres variables. Ainsi, il semble logique d'utiliser l'information contextuelle apportée par la détection des phases pour aider à la détection des étapes et des activités. Effectivement, si la probabilité est élevée d'être dans la phase ouverture, il y a peu de chances que l'étape en cours de réalisation soit la mise en place de l'implant (« implantation ») ou la suture (« hydrosuture » ou « suture fil ») par exemple. De même les connaissances sur les étapes probables apportent une information pour la reconnaissance des activités. A l'inverse les probabilités d'appartenance aux étapes ou aux activités peuvent apporter une information pertinente pour la détection des phases. Nous cherchons donc à construire un modèle statistique qui représente les relations entre les différents niveaux de granularité.

IV.1.2.3.1 Arbres de décision

Dans la littérature, Forestier et al. utilisent des **arbres de décision** pour déduire la phase la plus probable compte tenu de la séquences d'activités $A_w = \{a(t-w), \dots, a(t-1), a(t)\}$ [26]. L'activité au temps t est nommée $a(t)$ et w représente la taille de la fenêtre pour laquelle les activités sont considérées. Les arbres de décision permettent d'aider à la prise de décision à l'aide d'un modèle graphique de type « arbre ». Chaque nœud représente les différentes variables pouvant influencer la décision à prendre. Lors du parcours du graphe, un choix est fait à chaque nœud. A l'extrémité des branches (feuilles), les distributions des différents choix possibles sont données pour la décision à prendre en fonction des décisions prises à chaque étape. Forestier et al. combinent les densités de probabilités pdf_a des prédictions relatives aux activités présentes dans la fenêtre [26]. Ainsi, la probabilité de chaque phase est calculée de la manière suivante :

$$\hat{p}\,(p_j|A_w) = \sum_{a \in A_w} \text{pdf}_a(p_j)$$

Cette méthode permet de déduire la phase en fonction de l'activité en cours de réalisation et des précédentes (au sein d'une fenêtre de taille fixe). Dans notre cas, nous ne connaissons pas les activités, mais nous pouvons, à l'aide d'une méthode de CBVR connaître la probabilité de réalisation des différentes activités au temps t. Cette information peut donc être utilisée pour déduire la phase la plus probable. Néanmoins, nous souhaitons également utiliser notre connaissance de la phase la plus probable pour affiner la reconnaissance des activités. Cela n'est pas possible avec les arbres de décision, et il semble plus judicieux d'utiliser dans cette situation un réseau bayésien.

IV.1.2.3.2 Réseaux bayésiens

Les réseaux bayésiens sont des représentations graphiques de la causalité. Ils modélisent de façon intuitive l'influence d'un événement sur un autre [80]. Ils appartiennent à la famille des graphes acycliques orientés (DAG pour « Directed Acyclic Graph » en anglais) (cf paragraphe IV.1.1) et permettent de calculer des probabilités conditionnelles. Les nœuds parents représentent les causes et les nœuds fils les conséquences. Comme cela est démontré dans [80], bien que la flèche soit orientée d'un nœud A vers un nœud B, l'information peut circuler dans les deux sens

et dans une relation de cause à effet, une connaissance de l'effet peut également permettre de déduire des connaissances sur les causes. En d'autres termes :

> *« S'il existe une relation causale de A vers B, toute information sur A peut modifier la connaissance que j'ai de B, et, réciproquement, toute information sur B peut modifier la connaissance que j'ai de A. »* [80]

Les réseaux bayésiens s'appuient, pour modéliser les relations de cause à effet, sur le théorème de Bayes. Si un événement B s'est produit, la probabilité que celui-ci ait été produit par A est donnée par la relation suivante :

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Un réseau bayésien est alors défini par les relations de dépendances (ou d'indépendances conditionnelles) données par le graphe (description qualitative), et les probabilités conditionnelles associées à ces relations (description quantitative).

Ce type de modèles semble particulièrement pertinent pour modéliser les relations de cause à effet qui existent entre les différents niveaux de descriptions de la chirurgie. Ainsi, comme cela est présenté dans la Figure 46, les niveaux de granularité les plus fins représentent les causes et les plus élevés les conséquences.

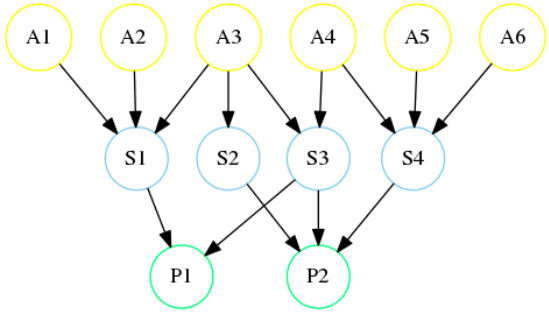


Figure 46. Exemple d'un DAG d'un réseau bayésien modélisant 3 niveaux de granularité ; les nœuds jaunes représentent les activités, les nœuds bleus les étapes et les nœuds verts les phases

Les réseaux bayésiens semblent donc plus adaptés à notre objectif que les arbres de décision du fait de leur capacité à propager l'information entre les différents niveaux, à la fois des causes vers les effets mais également des effets vers les causes.

IV.1.3 Synthèse

Il existe différentes méthodes pour apporter de l'information contextuelle afin d'aider au séquençage automatique d'une vidéo de chirurgie. L'apport d'information contextuelle passe par une modélisation du processus chirurgical, permettant de choisir la séquence de labels la plus

probable ayant produit la séquence d'observations. Cela se fait via la construction d'une chirurgie moyenne ou via l'utilisation d'un modèle statistique. Suite à l'étude des différentes méthodes qui existent dans la littérature, nous avons choisi d'évaluer les deux approches. La première approche consiste à construire une chirurgie moyenne. La méthode s'appuyant sur la construction d'une chirurgie moyenne unique via l'algorithme DTW n'étant pas adaptée à notre description de la chirurgie, nous avons choisi d'implémenter une approche originale de construction de chirurgie moyenne sous forme d'arbre, via une adaptation de la méthode proposée par Piciarelli et al. [44]. La seconde approche consiste à s'appuyer sur un modèle statistique tirant avantage de notre description multi-échelle de la chirurgie. Ainsi nous proposons un modèle statistique permettant à la fois de modéliser le déroulement temporel de la chirurgie aux différents niveaux de description, mais également d'utiliser les relations qui existent entre ces niveaux pour affiner la reconnaissance automatique des tâches chirurgicales. Les deux approches sont présentées dans les deux parties IV.2 et IV.3.

IV.2 Construction d'arbres (Piciarelli et al.)

Nous avons choisi, dans un premier temps, de tester une méthode s'appuyant sur la construction d'une chirurgie moyenne. La méthode basée sur l'algorithme DTW présente certaines limites face à notre problème (paragraphe IV.1.2.1). Notamment, la chirurgie moyenne modélise un enchaînement unique des phases chirurgicales. Or les descriptions en tâches, en étapes et en activités chirurgicales de la chirurgie de la cataracte ne permettent pas de décrire toutes les chirurgies avec le même enchaînement. En effet, comme cela a été mentionné dans le paragraphe II.1.1, une même tâche peut être réalisée plusieurs fois dans une même chirurgie si cela est nécessaire. De plus, de par la variété des chirurgiens représentés dans la base de données, l'enchaînement des tâches n'est pas toujours identique d'une chirurgie à l'autre. Il en est de même pour les étapes et les activités. Le processus chirurgical ne peut donc pas être représenté par une chirurgie moyenne unique. C'est pourquoi nous nous sommes orientés vers une méthode de construction d'arbre, adaptée de la méthode proposée par Piciarelli et al. [44] introduite dans le paragraphe III.1.2.2. Dans cette méthode, chaque nœud qui compose l'arbre est une moyenne de sections de vidéos de la base d'apprentissage. La structure de l'arbre représente les différents déroulements possibles de la chirurgie. Cela permet de prendre en compte différents cas d'ordonnancement des tâches chirurgicales. Nous avons donc décidé d'adapter cette méthode à l'analyse automatique des chirurgies de la cataracte. Pour cela, deux adaptations ont été proposées. La première consiste à réaliser un apprentissage supervisé, et à chercher à reconnaître automatiquement les cas anormaux (paragraphe IV.2.2.1). La seconde méthode mise en place consiste à utiliser les annotations des chirurgiens pour réaliser un apprentissage supervisé de l'arbre (paragraphe IV.2.2.2).

IV.2.1.1 Méthode de Piciarelli et al.

Piciarelli et al. cherchent à analyser automatiquement les trajectoires de véhicules dans des vidéos de télésurveillance [44]. Comme cela a déjà été introduit dans le paragraphe III.1.2.2, l'ensemble des trajectoires de la base d'apprentissage est partitionné et modélisé par un ensemble de nœuds organisés selon une structure en arbre. Chaque nœud représente un groupe de sections de trajectoires similaires. Le principe de partitionnement des trajectoires, introduit dans le paragraphe III.1.2.2, est résumé dans la Figure 47 et le processus de construction est présenté dans la Figure 48.

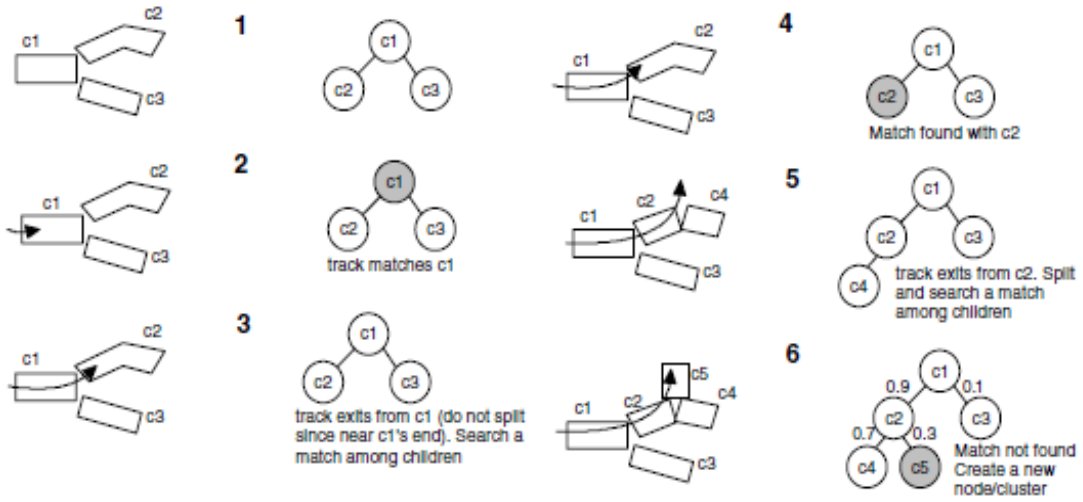


Figure 47. Mise à jour de l'arbre pour l'acquisition d'une nouvelle trajectoire [44]

Ce partitionnement est effectué « en direct » et de manière itérative, c'est-à-dire que chaque nouvelle trajectoire contribue à la construction du graphe. Plusieurs étapes sont également mises en place pour la maintenance du graphe. Tout d'abord, une étape de fusion des branches est effectuée. Pour cela, tous les nœuds d'un même niveau dans l'arbre (c'est-à-dire de même distance par rapport à la racine) sont comparés via la mesure de distance définie précédemment. Si deux nœuds sont suffisamment proches, alors ils sont fusionnés en un nouveau nœud. Ce dernier est alors une moyenne pondérée des deux anciens nœuds et hérite de l'ensemble de leurs nœuds fils. Suite à cette étape de fusion, il est possible qu'un nœud ait un seul nœud fils. Dans ce cas, ils sont concaténés en un nœud unique. Enfin, dans le cas ou une branche de l'arbre n'a pas été mise à jour depuis longtemps, celle-ci est supprimée.

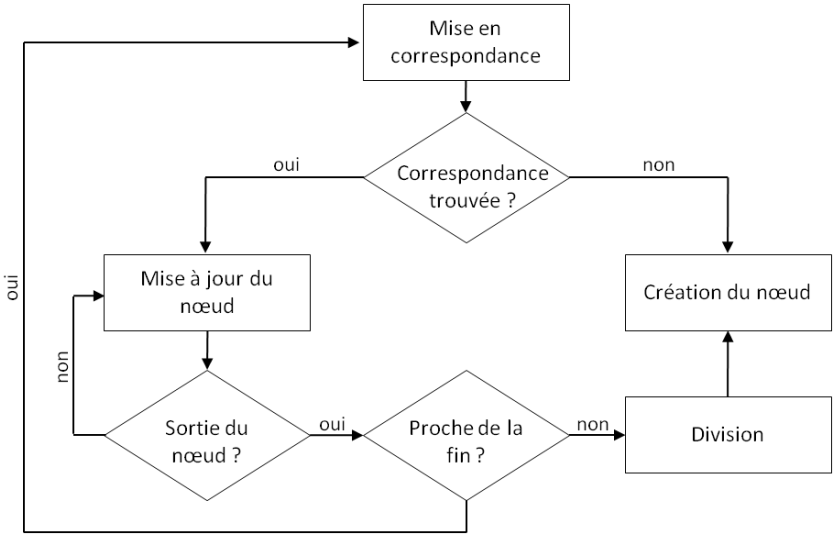


Figure 48. Processus de construction de l'arbre

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

A partir d'une centaine de trajectoires, le graphe est considéré comme suffisamment pertinent [44]. Toute nouvelle trajectoire va à nouveau participer à la construction du graphe, mais, de plus, le parcours de l'arbre qu'elle effectue va permettre de déduire des informations pertinentes sur cette trajectoire. Par exemple, un événement atypique pourra être détecté si elle traverse des nœuds de faibles probabilités ou qu'aucune correspondance n'est trouvée avec un nœud du modèle.

Cette méthode de modélisation des trajectoires possède de nombreux avantages et peut être adaptée dans notre cas à la modélisation du processus chirurgical, en remplaçant les trajectoires par les signatures visuelles des vidéos de chirurgie. Tout d'abord, elle est simple à mettre en place et apporte une alternative à la construction d'une chirurgie moyenne via l'algorithme DTW. Elle permet de modéliser différents déroulements possibles, sans se restreindre à un déroulement unique. De plus, elle structure l'espace de recherche. Il n'est alors pas nécessaire de comparer la vidéo en cours à tous les cas de la base, mais uniquement aux nœuds probables (c'est-à-dire les nœuds fils du nœud présent). Cela permet d'apporter une information contextuelle et d'accélérer les temps de calcul dans le cas de bases de données conséquentes. Il s'agit d'une méthode de construction a priori non supervisée. Elle permet de construire un modèle sans connaissance a priori de la base d'apprentissage. Les cas anormaux sont détectés lorsqu'aucune correspondance n'est trouvée avec un nœud du modèle : cela signifie que la vidéo en cours d'acquisition représente un cas non représenté dans les vidéos d'apprentissage. De plus, la méthode de construction du modèle est itérative, chaque nouveau cas participe à l'actualisation du graphe. Il est ainsi possible de mettre le modèle à jour aisément en cas d'apparition de nouvelles techniques, de nouvelles façons de procéder, ou de nouveaux cas.

IV.2.2 Construction de l'arbre

IV.2.2.1 Méthode non supervisée

Pour essayer de construire un modèle de déroulement du processus chirurgical, nous avons adapté la méthode de construction d'arbres développée pour l'analyse du comportement de véhicules proposée par Piciarelli et al. [44]. Les trajectoires ont été remplacées par les signatures des vidéos chirurgicales. Chaque nœud représente alors une séquence de chirurgie moyenne. Pour cela, la même méthode de construction et de mise à jour a été mise en place. Chaque vidéo $\mathcal{V}$ est constituée d'un ensemble d'images $\{v_1, v_2, \dots, v_n\}$ relevées à intervalles réguliers où $v_{ij} = \{x_1, x_2, \dots, x_l\}$ représente la signature de l'image j. Chaque nœud $\mathcal{C}$ est représenté par une séquence moyenne $\{c_1, c_2, \dots, c_m\}$ où $c_{ij} = \{x'_1, x'_2, \dots, x'_l, \sigma^2_{ij}\}$. Le processus de construction de l'arbre est identique à la méthode de Piciarelli et al. [44] présentée dans la Figure 48. Pour chaque nouvelle image v_i de la vidéo requête, si le point $c_j = \{x_1, \dots, x_l, \sigma^2\}$ est l'élément le plus proche de $v_i = \{\hat{x}_1, \dots, \hat{x}_l\}$ alors le nœud est mis à jour de la manière suivante :

$$\begin{cases} x'_k = (1 - \alpha)x'_k + \alpha\hat{x}_k \,, 0 < k \leq L \\ \sigma^2 = (1 - \alpha)\sigma^2 + \alpha d_{ij}^2 \end{cases}$$

où d_{ij} est la distance Bhattacharrya entre le point v_i et le point c_j. Chaque nœud est alors représenté par une séquence moyenne.

Cette approche est intéressante car elle ne nécessite pas d'annoter la base de données, étape qui requiert beaucoup de temps. Elle permet de reconnaître des évènements atypiques et éventuellement de fournir des exemples de vidéos qui ont suivi un parcours similaire. En revanche, elle semble nécessiter une base d'apprentissage conséquente (100 trajectoires dans le cas de l'analyse des trajectoires de véhicule sur une voie rapide [44]). Nous avons dans notre cas une base de données annotée. C'est pourquoi nous nous sommes intéressés à une méthode d'apprentissage supervisée, qui permet de construire l'arbre à partir de notre connaissance des labels assignés à chacune des images des vidéos de la base d'apprentissage.

IV.2.2.2 Méthode supervisée

Dans la base de données annotée, chacune des vidéos a été segmentée manuellement en tâches (paragraphe II.2.2.1), en étapes, en activités et en phases chirurgicales (paragraphe II.2). Nous connaissons, pour chacune des vidéos, les chronométrages de chacune des tâches (étapes, activité et phases) chirurgicales. Nous avons donc utilisé cette information pour réaliser un apprentissage supervisé d'un arbre. Pour évaluer la méthode, seule la description initiale en tâches a été utilisée et chaque nœud de l'arbre représente alors une tâche chirurgicale. Ce travail a été réalisé avec l'aide de Mme Esther Puyol Anton, que j'ai encadrée durant un stage de trois mois réalisé au sein de l'équipe GD2MP du LaTIM. Le processus de construction de l'arbre est présenté dans la Figure 49.

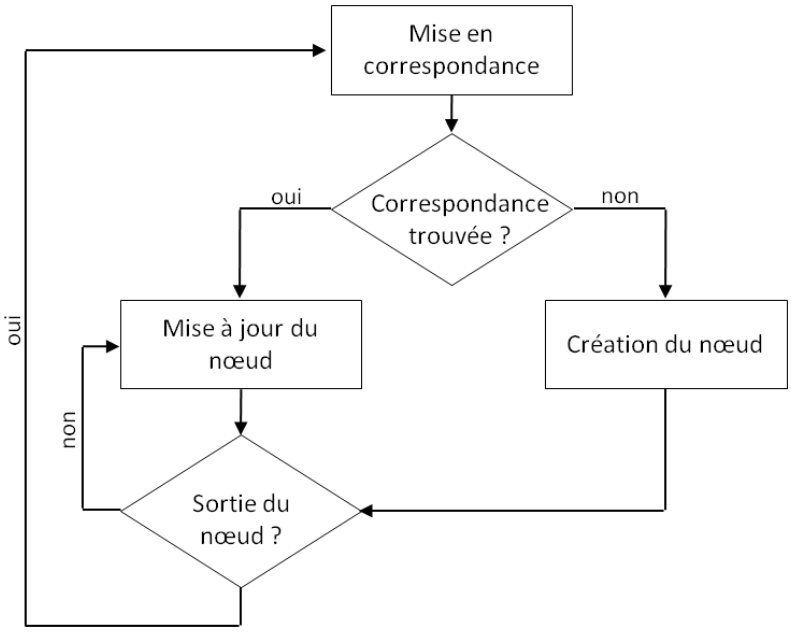


Figure 49. Méthode de construction supervisée d'arbres

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

Contrairement à la méthode précédente, l'étape de mise en correspondance n'est pas réalisée via la mesure de similitude, mais via les labels de chaque image. Soit $\mathcal{S} = \{S_1, S_2, \dots, S_n\}$ l'ensemble des labels possibles, à chaque image de la vidéo correspond un label S_i. Si le label de l'image en cours correspond au label d'un nœud, alors celui-ci est mis à jour tant que le label correspond. Lorsque le label change, une nouvelle étape de mise en correspondance est réalisée. Si aucune correspondance n'est trouvée, un nouveau nœud est créé.

IV.2.3 Inférence de l'arbre

La méthode d'inférence de l'arbre consiste à comparer la vidéo requête aux différents nœuds de l'arbre, en commençant par la racine de l'arbre. Le principe de parcours de l'arbre est présenté dans la Figure 50. Pour cela, les N premières images de la vidéo requête sont comparées aux séquences moyennes qui représentent chacune un nœud possible. La similitude entre les 100 premières images $\mathcal{V} = \{v_t \dots v_{t+N}\}$ de la requête et un nœud $\mathcal{C} = \{c_1 \dots c_m\}$ est calculée de la manière suivante :

$$D(\mathcal{T}, \mathcal{C}) = \frac{1}{N} \sum_{i=1}^{N} d(v_{t+i}, \mathcal{C})$$

avec

$$d(t_i, \mathcal{C}) = \min_j \left(\frac{d_{ij}}{\sqrt{\sigma_j^2}} \right), j \in \{[(1-\delta)i] \dots [(1+\delta)i]\}$$

Le nœud validé est celui pour lequel la mesure de similitude calculée est la plus faible. La mesure de similitude entre la requête et le nœud est évaluée pour chaque nouvelle image acquise. Lorsque cette distance dépasse un certain seuil, on sort du nœud et une correspondance est à nouveau cherchée parmi les nœuds fils (Figure 50).

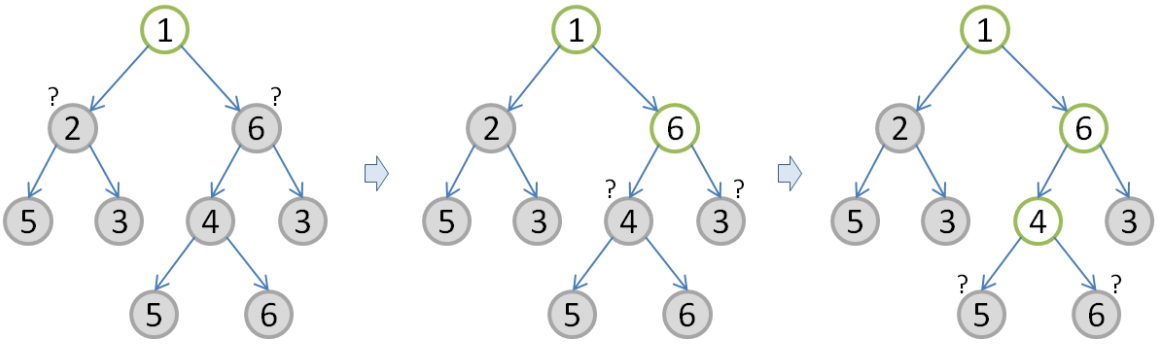


Figure 50. Parcours de l'arbre pour la reconnaissance automatique de tâches chirurgicales

IV.2.4 Résultats

Les deux méthodes d'apprentissage ont été évaluées avec les deux types de signatures présentées dans le paragraphe III.2.1.1 (page 51) : les histogrammes de mouvement et les histogrammes de mots visuels. La valeur du facteur d'agrandissement de la fenêtre glissante a été fixé à $\delta = 0{,}1$, comme pour l'évaluation de la mesure de similitude (paragraphe III.4.3.1, page 87).

IV.2.4.1 Méthode non supervisée

Dans le cas de la méthode d'apprentissage non-supervisée, il est difficile de distinguer le moment ou deux vidéos empruntent des « chemins » différents. En effet, comme le montre la Figure 51, il n'y a pas d'accroissement significatif de la distance, quelle que soit l'évolution de la chirurgie. On constate que les valeurs maximales de distance sont obtenues, pour les différents exemples au début de la vidéo. Cela peut s'expliquer par le fait que les vidéos sont très différentes en début de chirurgie, avant même que les gestes chirurgicaux ne débutent. Certaines actions, non-pertinentes d'un point de vue chirurgical, mais ayant un fort impact visuel, ont lieu pendant les premières secondes ou minutes. Par exemple, l'éclairage n'est pas toujours en route au démarrage de l'enregistrement, ou dans certains cas la référence de l'implant est placée dans le champ de vue de la caméra. Une autre explication vient du fait que la taille de la fenêtre glissante est très faible en début de chirurgie (elle grandit ensuite au cours du temps). Les correspondances possibles entre l'image en cours de la vidéo requête et les images de la vidéo de référence sont donc limitées. Les divergences observées en fin de vidéo (pour le graphe en haut à gauche et le graphe en bas à gauche), s'expliquent par le fait que la fenêtre glissante a dépassé la fin de la vidéo de référence. Dans ce cas, les mesures de distance sont mises à leur valeur maximale.

Il n'a donc pas été possible de construire un arbre. Contrairement aux trajectoires des véhicules, les signatures de nos vidéos sont beaucoup plus complexes que les trajectoires de véhicules. Les trajectoires de véhicules ont une forte reproductibilité, ce qui n'est pas le cas des signatures visuelles. Il est donc complexe de construire un arbre selon cette méthode. Néanmoins, le concept reste intéressant, et une autre piste consiste à utiliser les informations issues de la description de la base d'apprentissage pour construire l'arbre.

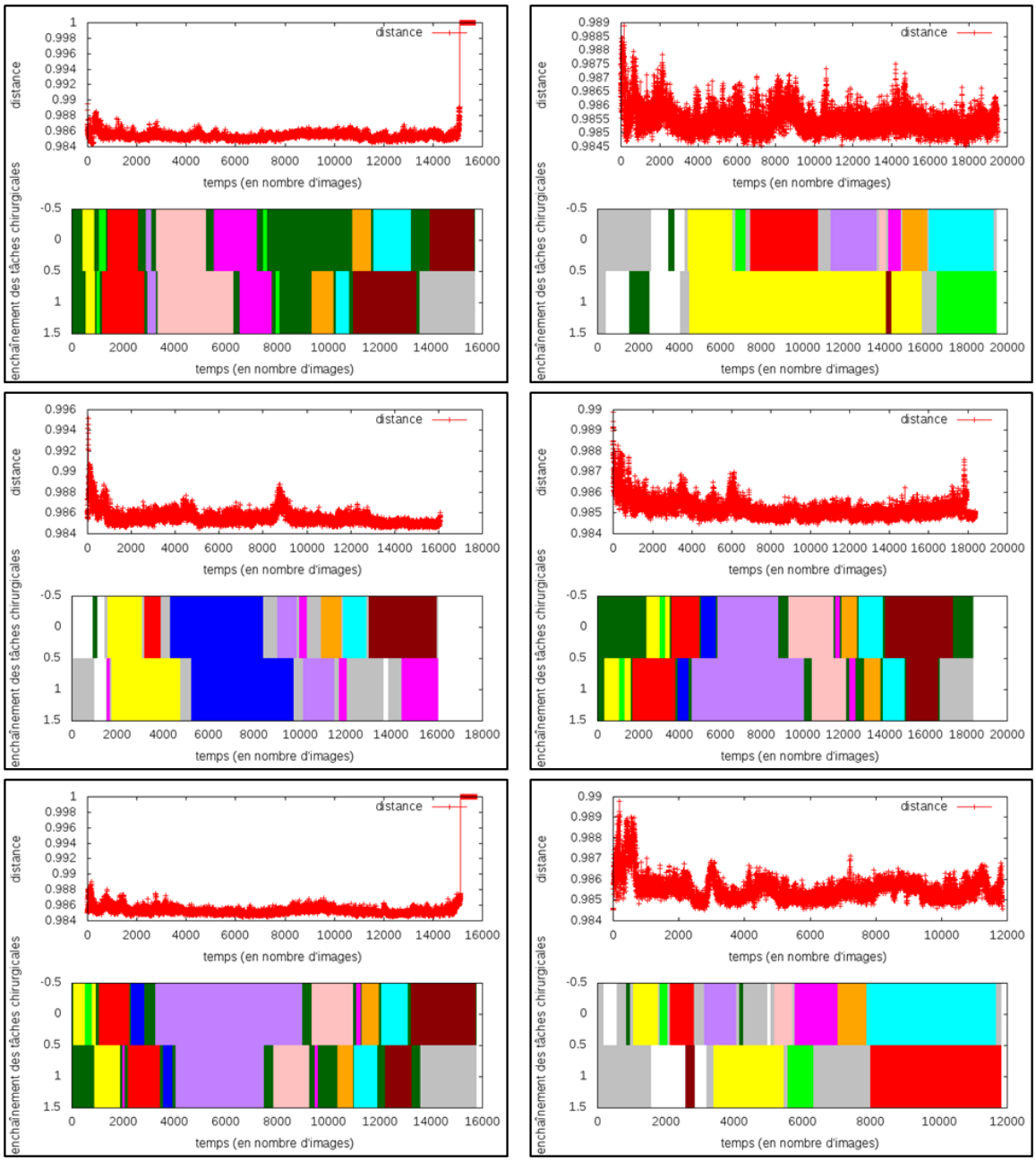


Figure 51. Exemples de comparaisons de vidéos 2 à 2 ; en haut, l'évolution de la mesure de distance entre les deux vidéos ; en bas, la comparaison de l'enchaînement des phases chirurgicales entre la vidéo requête (en haut) et la vidéo de référence (en bas)

IV.2.4.2 Méthode supervisée

Dans le cas de la construction supervisée de l'arbre, les différentes structures d'arbres obtenues sont présentées dans la Figure 52 et l'Annexe 1. Ces arbres ont été construits avec respectivement 10 et 42 vidéos, afin d'être plus lisibles. Pour la suite de l'évaluation, un arbre construit avec 161 vidéos a été utilisé. On peut constater que malgré le caractère reproductif de la chirurgie de la cataracte, il existe néanmoins une variabilité importante dans l'enchaînement des tâches chirurgicales. Ceci entraîne une démultiplication du nombre de branches dans l'arbre. La méthode a été évaluée avec les deux types de signatures présentées dans le paragraphe III.2.1.1 : les histogrammes de mouvement et les histogrammes de mots visuels. Le seuil a été fixé à une valeur de 2σ. Les résultats en termes de sensibilité et spécificité obtenus pour deux vidéos de la base de test sont présentés dans le Tableau 10.

Tableau 10. Résultats obtenus pour deux vidéos de test après inférence de l'arbre

	Vidéo 1		Vidéo 2	
	Sensibilité	Spécificité	Sensibilité	Spécificité
MH	0.003	0.710	0.012	0.799
BoW	0.028	0.786	0.008	0.696

Tout comme pour la méthode non supervisée, les résultats n'ont pas été probants du fait qu'il est difficile de déterminer les transitions d'un nœud vers un autre. Ainsi, les valeurs de sensibilité obtenues sont quasi nulles. De même que pour la méthode précédente, il est difficile de déterminer un seuil. En plus du problème de variabilité des signatures, déjà évoqué dans le paragraphe IV.2.4.1, se pose également le problème de la démultiplication des branches de l'arbre. Cela implique que dans une majorité de cas, peu d'exemples ont participé à la création de chacun des nœuds.

Pour pallier cela, nous avons évalué une évolution de la méthode, pour laquelle toutes les séquences représentant une même tâche participent à la construction d'une signature globale pour tous les nœuds représentant cette tâche. Ainsi, tous les nœuds représentant une même tâche chirurgicale auront la même signature moyenne. Cette approche a l'avantage d'augmenter le nombre d'exemples par nœud, mais fait disparaitre certaines spécificités. Par exemple, une incision réalisée juste avant la mise en place de l'implant pour agrandir l'incision déjà présente est un peu différente de l'incision réalisée en début de chirurgie. Or elles seront représentées par une même signature moyenne. La méthode a été évaluée uniquement avec la caractérisation des vidéos par des histogrammes de mots visuels. Le seuil a également été fixé à une valeur de 2σ. Les résultats, en termes de sensibilité et spécificité, obtenus pour les deux vidéos de test sont présentés dans le Tableau 11.

Tableau 11. Résultats obtenus pour deux vidéos de test après inférence de l'arbre (signatures globales)

	Vidéo 1		Vidéo 2	
	Sensibilité	Spécificité	Sensibilité	Spécificité
BoW	0.007	0.847	0.009	0.920

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

Cette approche n'a également pas permis d'apporter des résultats satisfaisants, avec des valeurs de sensibilités obtenues très faibles.

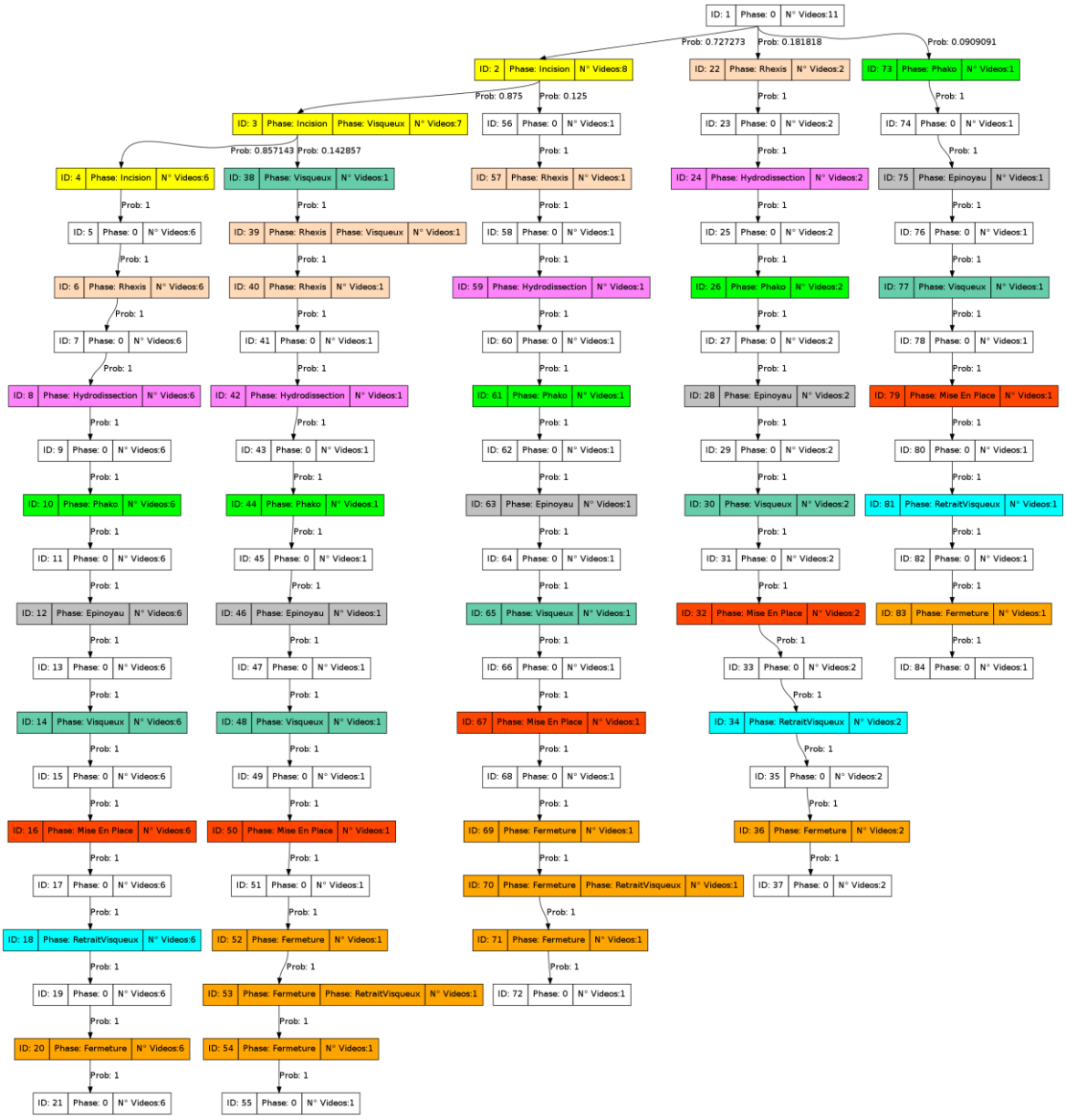


Figure 52. Exemple d'arbre obtenu avec 11 vidéos

IV.2.5 Synthèse

Cette approche de construction d'arbres, bien qu'intéressante car intuitive et simple à mettre en œuvre n'a pas donné les résultats escomptés. Cela est probablement lié au fait que contraire-

ment aux trajectoires des véhicules, les signatures sont plus variables. Il est donc difficile de distinguer une divergence entre le nœud qui est en train d'être parcouru et la vidéo requête. Il n'a donc pas été possible de construire automatiquement un arbre selon la méthode de Piciarelli et al. [44] en remplaçant les trajectoires par les signatures des vidéos. Le problème de l'apprentissage de la structure de l'arbre a été résolu en utilisant notre connaissance de la base de données. La construction de l'arbre a été contrainte par la connaissance des labels (tâches chirurgicales) associés à chacune des images. Ainsi, chaque nœud de l'arbre représente une tâche chirurgicale. Cette approche a permis de constater que malgré le caractère reproductif de la chirurgie de la cataracte, le nombre de branches dans l'arbre est très élevé. On a vu de plus dans le paragraphe II.4 que l'enchaînement des étapes et des activités est encore plus variable que celui des tâches. Cela devrait engendrer un nombre de branches encore plus élevé pour ces niveaux de description. La méthode d'inférence de l'arbre n'a pas permis de discerner les transitions entre les tâches chirurgicales. De la même manière que pour la méthode d'apprentissage non supervisée, il est difficile de distinguer une divergence entre le nœud et la vidéo requête. Pour pallier ce problème, il pourrait être judicieux de travailler avec des sous-séquences de tailles fixes et non plus image par image. Il pourrait également être intéressant de créer des nœuds qui regroupent un ensemble de sous-séquences et non pas des nœuds représentés par une séquence moyenne. La comparaison d'une vidéo requête avec le modèle se ferait alors par une recherche des sous-séquences les plus proches parmi les groupes de sous-séquences (nœuds) possibles.

Les résultats n'étant pas prometteurs pour un niveau de granularité élevé (les tâches chirurgicales), nous n'avons pas poursuivi dans cette voie. En effet, le déroulement de la chirurgie à des niveaux de granularité plus fins (étapes et activités) est encore plus complexe (paragraphe II.4.3 et II.4.4). Nous sommes donc passés à la seconde approche, à base de graphes, qui tire avantage de notre description multi-échelle de la chirurgie.

IV.3 Modélisation statistique multi-échelles

La réflexion menée sur la description du processus chirurgical a abouti à la création d'une nouvelle description de la chirurgie de la cataracte à différents niveaux de granularité. Nous avons choisi d'utiliser cet aspect multi-échelles pour construire notre modèle statistique. L'objectif de cette modélisation est de pouvoir séquencer automatiquement une vidéo de chirurgie requête à plusieurs niveaux de précision en utilisant les avantages des différents niveaux. Pour cela, nous avons choisi de nous orienter vers un modèle combinant un réseau bayésien et des modèles markoviens ou des CRF (Figure 53). Le réseau bayésien modélise les relations de cause à effet qui existent entre les niveaux. Les HMM et les CRF peuvent être utilisées indifféremment pour modéliser les déroulements temporels de la chirurgie pour chaque description. Par souci de simplification, la méthode sera présentée dans le cas des HMM, mais des résultats seront également donnés dans le cas des CRF.

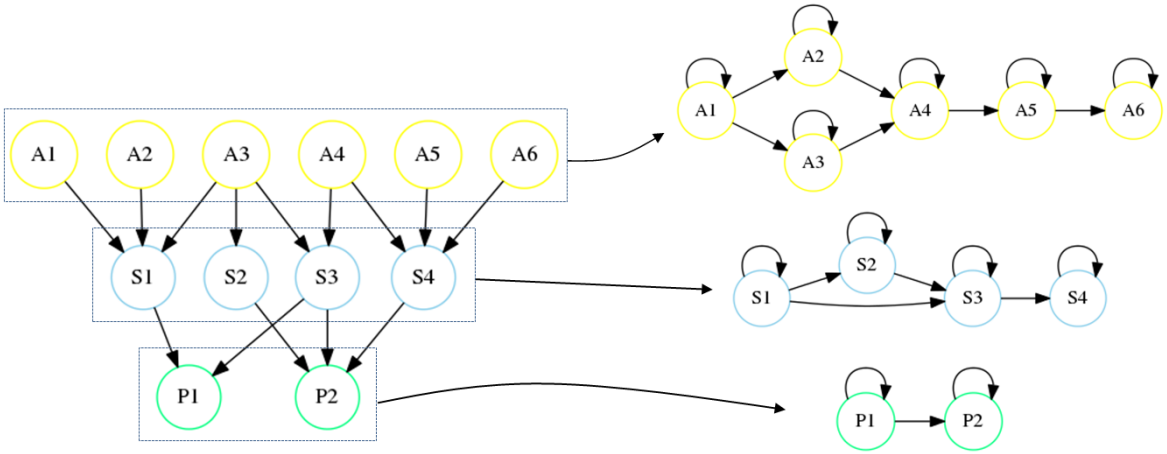


Figure 53. Principe général de la modélisation multi-échelles de la chirurgie, combinant un réseau Bayésien (à gauche) et 3 modèles de Markov (à droite)

Des observations sont générées à pas de temps fixes lors de l'acquisition de la nouvelle vidéo. Deux types d'observations seront évalués dans cette partie : la présence des instruments chirurgicaux dans le champ de vu de la caméra et l'analyse du mouvement dans la vidéo associé à une méthode de CBVR. Ces observations fournissent des preuves au réseau bayésien qui va émettre des probabilités de réalisation pour chacun des labels possibles (aux différents niveaux de description). Ces probabilités de réalisation sont alors utilisées en entrée des différents HMM pour déduire les labels les plus probables pour chacune des images. Le modèle a été construit et évalué avec deux niveaux de description pour déterminer la faisabilité de la méthode. Ainsi, les descriptions en étapes et en phases chirurgicales ont été utilisées pour les différents tests.

IV.3.1 Construction du modèle

Le modèle statistique permet de calculer les probabilités de réalisation des différentes étapes et phases chirurgicales à partir des observations obtenues pour chaque sous-séquence. Ce modèle

est composé d'un réseau bayésien et de HMM. Le réseau bayésien modélise les relations qui existent entre les différents niveaux de granularité (observations, étapes et phases). Les modèles de Markov quand à eux modélisent le déroulement temporel de la chirurgie de la cataracte à chacun des niveaux de descriptions (étapes et phases). Les différents blocs du modèle sont appris à partir de la base de données, afin de permettre ensuite de déduire, par inférence, les labels les plus probables pour chacune des images qui composent la vidéo. La labellisation peut ensuite être effectuée pour chaque niveau de description en choisissant le label qui a la plus grande probabilité de réalisation. Ainsi, la vidéo est séquencée temporellement aux différents niveaux de description. La méthodologie du système est présentée dans la Figure 54.

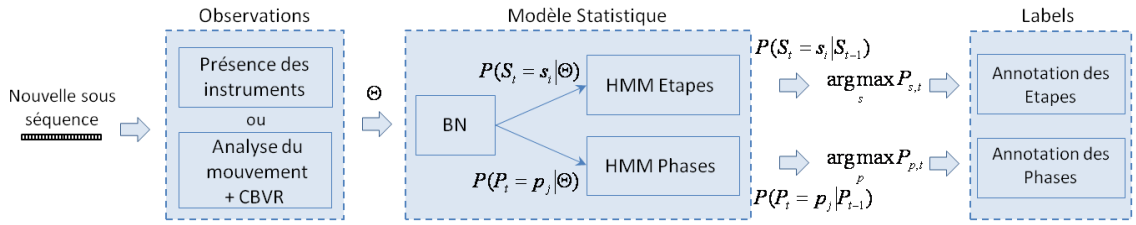


Figure 54. Méthodologie du système d'annotation automatique

La construction du modèle se fait par apprentissage des différentes structures qui le composent et des probabilités associées. Chaque vidéo de la base d'apprentissage est découpée en sous-séquences de taille fixe, comme présenté dans le paragraphe IV.3.2.1. A chaque sous-séquence de la base d'apprentissage est associé un label. Ce label est celui de la première image de la sous-séquence. Les relations entre les labels de chaque niveau de description pour une sous-séquence et les transitions observées d'une sous-séquence vers la suivante vont permettre de construire les structures et les probabilités des différents blocs du modèle. L'apprentissage des différents blocs du modèle est présenté dans les paragraphes suivants.

IV.3.1.1 Réseau Bayésien

Comme nous l'avons vu dans le paragraphe IV.1.2.3.2, un réseau bayésien est composé à la fois d'un graphe et d'un ensemble de probabilités conditionnelles. L'apprentissage se fait alors selon deux aspects : l'apprentissage de la structure et l'apprentissage des paramètres, c'est-à-dire des probabilités conditionnelles associées à chaque arc.

Nous cherchons à modéliser les influences entre les différents labels, à savoir les probabilités de cooccurrence d'une phase donnée et d'une étape donnée (ou d'une étape donnée et d'une activité donnée). Les variables représentées dans le réseau bayésien seront donc associées à ces labels. Les N_s labels des étapes qui ont ainsi été rencontrés dans la base d'apprentissage forment l'ensemble $\mathcal{S} = \{(s_k)_{0\leq k<N_s}\}$. Lorsque deux étapes peuvent se dérouler en même temps, une variable supplémentaire, dont le label est composé des noms des deux étapes, est créée. De même, Les N_p labels des phases qui ont été rencontrées dans la base d'apprentissage forment l'ensemble $\mathcal{P} = \{(p_j)_{0\leq j<N_p}\}$. Rappelons que deux phases ne peuvent se chevaucher temporellement. Notons que ces ensembles constituent également les états des HMM.

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

Nous cherchons, lors de l'analyse d'une vidéo, à déterminer les probabilités de réalisation des différentes variables (labels), à différents instants de la chirurgie, en fonction d'observations issues de notre analyse du flux vidéo. A priori, ce lien entre le flux vidéo et le déroulement de la chirurgie (étapes et phases) peut se faire soit au niveau du réseau bayésien, soit au niveau des HMM, soit les deux. Mais puisque le réseau bayésien fournit les entrées des HMM, il nous semble plus efficace de faire ce lien au niveau des réseaux bayésiens. Pour cela, nous intégrons au réseau bayésien des variables supplémentaires, que nous appellerons « variables d'observation », et qui vont recevoir à chaque pas de temps des observations issues de l'analyse du flux vidéo. Notons que l'introduction de variables d'observations tient à une limitation des réseaux bayésiens : les messages que reçoit un réseau bayésien pour mettre à jour son état sont des messages binaires, appelés « preuves ». Une preuve indique avec certitude que l'une des variables du réseau a pris telle ou telle valeur. Or notre analyse du flux vidéo ne permet en aucun cas d'affirmer avec certitude si telle ou telle étape ou phase est actuellement en cours de réalisation (ce chapitre serait alors superflu). Par contre, elle nous permet d'affirmer avec certitude que l'intensité moyenne des pixels est supérieure à un seuil τ donné ou qu'au moins n des k sous-séquences les plus proches de la sous-séquence courante proviennent d'une étape donnée ou d'une phase donnée. Rien ne nous empêche donc de créer un « nœud d'observation » représentant par exemple l'affirmation « l'intensité moyenne des pixels est supérieure à τ ». Aux deux ensembles de nœuds précédemment décris, traduisant l'état d'avancement de la chirurgie, s'ajoute donc un troisième ensemble permettant de faire le lien entre le flux vidéo et l'état d'avancement : chacun de ces nœuds est appelé « nœud d'observation ». Les nœuds $\mathcal{V} = \{S_1, \dots, S_N\}_{N=N_s+N_p+N_o}$ du graphe représentent alors l'ensemble des labels possibles $\mathcal{S} = \left\{(s_k)_{0\leq k<N_s}\right\}$ et $\mathcal{P} = \left\{(p_j)_{0\leq j<N_p}\right\}$, ainsi que l'ensemble des observations $\mathcal{O} = \left\{(o_l)_{0\leq l<N_o}\right\}$. Chaque nœud peut prendre deux valeurs : $x_1 =$ « réalisé » ou $x_0 =$ « non réalisé ». Pour chaque nouvelle inférence (à chaque nouvelle observation), les probabilités de réalisation de chacun des labels pourront alors être calculées.

L'apprentissage de la structure du réseau bayésien a été réalisée par comptage des cooccurrences entre les différents niveaux de granularité. Dans notre cas, la construction du réseau bayésien est simplifiée par l'existence des trois niveaux de granularité : les observations (qui peuvent être vues comme le niveau de granularité le plus fin), les étapes et les phases. En effet, la création de variables traduisant le chevauchement temporel (d'étapes) rend inutile les liens de cooccurrence entre nœuds du même niveau Les arcs du réseau traduisent les liens de cause à effet entre ces niveaux. La définition des trois niveaux impose l'orientation des arcs, des causes vers les effets, c'est-à-dire des observations vers les étapes et des étapes vers les phases (Figure 59). Cela impose également un certain nombre de relations d'indépendances, en limitant les possibilités de relation entre variables. Cette structure a également l'avantage d'exclure toute possibilité de cycles. Il reste à déterminer l'existence des arcs entre deux nœuds de niveaux successifs. Pour cela, on observe les cooccurrences qui existent pour chaque sous-séquence de la base d'apprentissage. Pour chaque sous-séquence, une observation est générée. On connaît, grâce aux annotations des chirurgiens, les labels « étape » et « phase » associés à cette observation. Ainsi, à chaque sous-séquence est associé un triplet $< o, s, p >$, $o \in \mathcal{O}, s \in \mathcal{S}$ et $p \in \mathcal{P}$. Si pour un tel triplet, l'arc est inexistant entre o et s, alors celui-ci est créé, il en est de même pour l'arc (s, p).

L'**apprentissage des paramètres** du réseau bayésien revient à appendre les probabilités conditionnelles associées à chaque nœud. Pour cela nous nous appuyons sur une approche statistique

classique : la méthode du maximum de vraisemblance. Les nœuds parents peuvent prendre différentes configurations possibles qui correspondent aux différentes combinaisons (vrai ou faux) possibles du nœud parents. Ainsi, s'il existe r nœuds parent, chaque nœud pouvant prendre 2 valeurs possibles, il existe alors 2^r combinaisons possibles. La probabilité que le nœud V_i prenne la valeur x_k sachant que ses parents sont dans la configuration w_j est donnée par la relation suivante :

$$P\big(V_i = x_k | \mathrm{Pa}(V_i) = w_j\big) = \frac{N_{ijk}}{\sum_k N_{ijk}}$$

où N_{ijk} représente le nombre d'instances dans la base où $V_i = x_k$ et que ses parents $\mathrm{Pa}(V_i)$ sont dans la configuration w_j. La probabilité de réalisation de la phase ou de l'étape chirurgicale, sachant la combinaison de ces parents, est évaluée pour chaque nœud.

Un **élagage** de la structure du graphe est ensuite nécessaire une fois les probabilités conditionnelles apprises. Les graphes obtenus peuvent être très complexes avec un grand nombre de relations. L'augmentation du nombre de parents augmente le nombre de combinaisons possibles de façon exponentielle, ce qui implique un grand nombre de probabilités conditionnelles à stocker. De plus, plus le graphe est complexe plus son parcours est long. Afin de restreindre la complexité de la structure du graphe, les arcs les moins pertinents ont été éliminés. Nous avons tout d'abord utilisé le calcul de l'information mutuelle comme critère de pertinence. L'information mutuelle est très souvent utilisée lors de la construction des réseaux bayésiens, via la maximisation d'une fonction de score, qui s'appuie sur le calcul l'information mutuelle entre un nœud et ses parents [81]. Etant donnée la distribution de probabilité jointe $P(x, y)$ pour le couple de variables aléatoires (X, Y), l'information mutuelle entre ces deux variables peut être évaluée de la manière suivante :

$$\mathrm{MI}(X, Y) = \sum_{x,y} P(x, y) \log\left(\frac{P(x, y)}{P(x)P(y)}\right)$$

Dans notre cas, nous évaluons la pertinence de chaque arc (V_j, V_i) en calculant l'information mutuelle entre un nœud V_i et son parent V_j via la formule suivante [80] :

$$\mathrm{MI}(V_i, V_j) = \frac{1}{N}\sum_{j=1}^{2}\sum_{k=1}^{2} N_{ik} \log\frac{N N_{ijk}}{N_{ik} N_{ij}}$$

Nous avons également évalué un autre critère, qui consiste à supprimer les arcs représentant un faible nombre de cooccurrences. Ce critère nous semble plus adapté dans notre cas. En effet, l'utilisation de l'information mutuelle permet de représenter de la meilleure façon possible les relations au sein du réseau bayésien, en supposant que les probabilités conditionnelles sont fiables. Mais dans notre cas, les probabilités conditionnelles sont obtenues par comptage des cooccurrences dans la base de données (une base de données de 30 vidéos) : elles reposent donc sur l'hypothèse « fréquence = probabilité », qui est mise à mal lorsque le nombre d'exemples d'apprentissage est limité. L'élimination des arcs représentant un faible nombre de cooccurrences permet d'éliminer les cas faiblement représentés, pouvant résulter d'une erreur d'annotation.

IV.3.1.2 HMM

Les modèles de Markov modélisent le déroulement temporel du processus chirurgical, via un modèle graphique. Celui-ci représente les transitions possibles entre les différents labels. Les états des deux HMM sont l'ensemble des labels possibles à chaque niveau, respectivement $\mathcal{S} = \{(s_k)_{0\leq k<N_s}\}$ et $\mathcal{P} = \{(p_j)_{0\leq j<N_p}\}$. Un modèle de Markov est construit par niveaux de description (étapes et phases). L'apprentissage se fait par comptage des transitions de l'état S_i vers de l'état S_j dans la base d'apprentissage. L'observation des transitions se fait à pas de temps régulier. On notera hmm_{s} et hmm_{p} les pas de temps utilisés respectivement pour le HMM des étapes et pour celui des phases.

L'apprentissage des HMM est relativement simple et consiste à construire la matrice $\boldsymbol{A}$ des transitions et le vecteur $\boldsymbol{\pi}$ des probabilités initiales. La matrice de transition $\boldsymbol{A} = (a_{ij})_{0\leq i,j<N}$ est définie de la manière suivante :

$$a_{ij} = P(q_t = S_j | q_{t-1} = S_i)$$

avec $a_{ij} \geq 0$ et $\sum_{j=1}^{N} a_{ij} = 1$. Chaque élément de la matrice a_{ij} de la matrice représente la probabilité de transition vers l'état S_j, sachant que l'on est dans l'état S_i. L'apprentissage de la matrice $\boldsymbol{A}$ se déduit directement de l'apprentissage de la structure et $a_{ij} = \frac{N_{ij}}{\sum_j N_{ij}}$ ou N_{ij} représente le nombre de cas dans la base d'apprentissage où l'on observe une transition de l'état S_i vers de l'état S_j. La matrice d'observation $\boldsymbol{B}$, est définie par ces éléments $b_i(k)$, où $b_i(k) = P(o_t = \mathrm{k} | q_t = S_i)$ la probabilité d'émettre le symbole k en étant dans l'état S_i. Elle permet de déduire, à partir des observations la probabilité que celle-ci est été générée par l'état S_i. Dans notre cas, il n'est pas nécessaire de connaître la matrice d'observation $\boldsymbol{B}$ car la probabilité $P(s_k|o_t)$ d'obtenir le label s_k sachant l'observation o_i est obtenue via le réseau bayésien. En effet, à chaque pas de temps hmm_{s} et hmm_{p}, l'inférence du réseau bayésien est réalisée, en fonction des observations obtenues. Cette inférence permet d'évaluer, en fonction des différentes preuves les probabilités de réalisation de chacun des labels (paragraphe IV.3.3.2). Ces probabilités seront utilisées par l'algorithme de Viterbi [78] pour l'inférence des HMM (paragraphe IV.3.3.2.2.

IV.3.1.1 CRF

Le modèle a également été évalué en remplaçant les HMM par des CRF (Figure 55). Les CRF semblent donner de meilleurs résultats que les HMM dans le cadre de l'analyse de vidéos chirurgicales [30][51]. Soit $\mathcal{S} = \{(s_k)_{0\leq k<N_s}\}$ (ou $\mathcal{P} = \{(p_j)_{0\leq j<N_p}\}$) l'ensemble des labels possibles et $\mathcal{Y} = \{y_t\}_{0<t\leq T}$ une séquence de labels avec $y_t \in \mathcal{S}$ (ou $y_t \in \mathcal{P}$). L'objectif est de labelliser une nouvelle observation $\boldsymbol{o}_t$ en choisissant le label y_t qui maximise la probabilité $P(y_t|\boldsymbol{o}_t)$. Les CRF dépendent de potentiels (ou fonctions caractéristiques). Attention à ne pas confondre les observations du CRF, qui proviennent du réseau bayésien, et les observations du réseau bayésien, qui proviennent de l'analyse des vidéos.

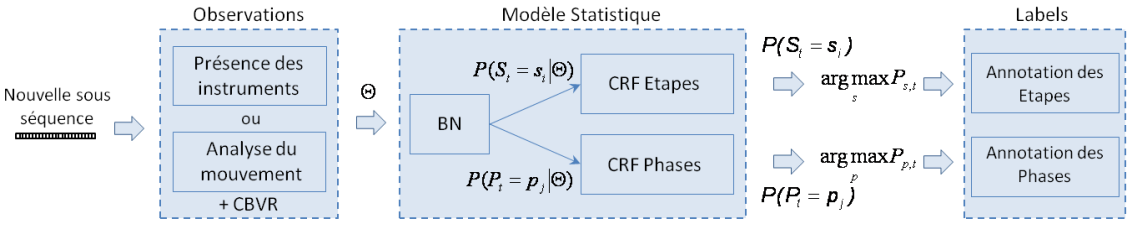


Figure 55. Méthodologie du système d'annotation automatique pour lequel les HMM ont été remplacés par des CRF pour la modélisation du déroulement temporel de la chirurgie

Les potentiels « unaires » $\psi_t^u(y_t, \boldsymbol{o}_t)$ représentent les scores d'assignation d'un label y_t la sous-séquence v_t. Dans notre cas, la définition des potentiels ne peut que dépendre de des évènements présents et passés. L'ensemble des potentiels définis dans notre cas dépend du contenu visuel des sous-séquences et est obtenu via le réseau bayésien. Le réseau bayésien fourni pour chaque sous-séquence v_t les probabilités de réalisation de chacune des étapes et des phases chirurgicales $P(s_k|o_t)$ et $P(p_k|o_t)$. L'ensemble de potentiels est construit à partir des probabilités obtenues l'ensemble de sous-séquences $\{v_{t-l\Delta_s}\}_{0\leq l\Delta_s\leq L}$. Les valeurs L et Δ_s sont calculées par apprentissage. Les potentiels « unaires » sont alors définis de la manière suivante :

$$\psi_t^u(s_t, \boldsymbol{o}_t) = \log\left(P(s_k|\boldsymbol{o}_{t-l\Delta_s})\right), 0 \leq l\Delta_s \leq L$$

Les potentiels « binaires » $\psi_{t-1,t}^b(y_{t-1}, y_t, \boldsymbol{o}_t)$, avec $b = 1 \ldots \frac{N(N-1)}{2}$ (avec $N = N_p$ ou N_s) , sont les potentiels représentant la pertinence de passer d'un label y_{t-1} au label y_t lorsque l'on passe de la sous-séquence v_{t-1} à la sous-séquence v_t. La relation entre deux sous-séquences est donnée par la probabilité $P_{s_i,s_j} = \frac{N_{ij}}{\sum_j N_{ij}}$, calculée à partir de la base de données. N_{ij} représente le nombre de cas dans la base d'apprentissage où l'on observe une transition de l'état S_i vers de l'état S_j. Les potentiels « binaires » sont alors définis de la manière suivante :

$$\psi_{t-1,t}^b(y_{t-1}, y_t, \boldsymbol{o}_t) = \log(P_{y_{t-1},y_t})$$

La librairie Wapiti[6], a été utilisée pour l'apprentissage et l'inférence des CRF et l'algorithme Quasi-Newton L-BFGS (« Limited-memory Broyden-Fletcher-Goldfarb-Shanno ») a été utilisé pour apprendre les poids $\boldsymbol{\lambda}$ et $\boldsymbol{\mu}$ qui permettent de pondérer les fonctions caractéristiques ψ_t^u et $\psi_{t-1,t}^b$ [82]. De plus, la librairie Wapiti6, utilisée pour l'apprentissage et l'inférence des CRF a été développée pour des méthodes d'analyse de texte et certaines adaptations ont dû être effectuées. La librairie lit les observations enregistrées comme des chaînes de caractères et non comme des nombres flottants. Or nos observations sont les probabilités de réalisation de chacun des labels. Deux probabilités proches, bien que similaires seront alors considérées comme des observations différentes. Il a donc été nécessaire de regrouper les observations en un nombre de classes fini et adéquat. Deux probabilités similaires appartiendront alors à une même classe et seront considérées par le système comme une même observation. Le nombre de classes choisi doit donc être

[6] https://wapiti.limsi.fr/

suffisamment grand pour différencier les observations, mais si le nombre de classes est trop élevé, il y aura alors trop peu d'exemples par observation.

IV.3.1.2 Retour du HMM vers le réseau bayésien

Une faiblesse de la méthode est que la connaissance issue de l'inférence des HMM (ou des CRF) n'est pas retransmise au réseau bayésien. Les différents éléments du modèle ne travaillent donc pas complètement ensemble. Pour remédier à cela, nous avons évalué une évolution du modèle pour laquelle les probabilités de réalisation de chacun des labels au niveau de description « phases », obtenues suite à l'inférence du HMM, servent également de preuves pour le réseau bayésien. Ces probabilités sont transmises au réseau bayésien via des nœuds d'observation raccordés aux nœuds des labels « phases » (Figure 56).

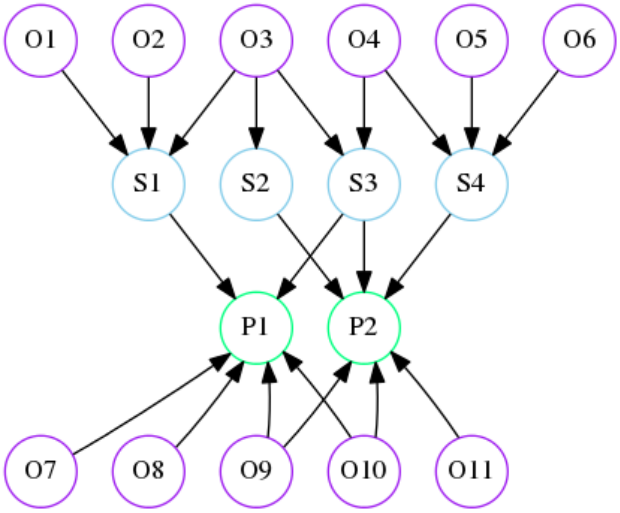


Figure 56. Exemple de réseau bayésien à deux niveaux de description, en violet les nœuds d'observation ; Les nœuds O7 à O11 permettent d'apporter les informations obtenus suite à l'inférence du HMM

Seuls les résultats du HMM phases sont retransmis au réseau bayésien car le HMM à ce niveau de description a permis d'améliorer de façon importante les performances de reconnaissance (paragraphe IV.3.3.3). Les résultats de l'utilisation du HMM étapes étant plus mitigés, nous n'avons pas jugé utile de complexifier le graphe en ajoutant d'autres nœuds à ce niveau de description. Ainsi, pour chaque nouvelle inférence du HMM, les informations sont transmises au réseau bayésien et apporteront ainsi des preuves supplémentaires lors de la labellisation de la sous-séquence suivante. La méthodologie de cette version du modèle est présentée dans la Figure 57.

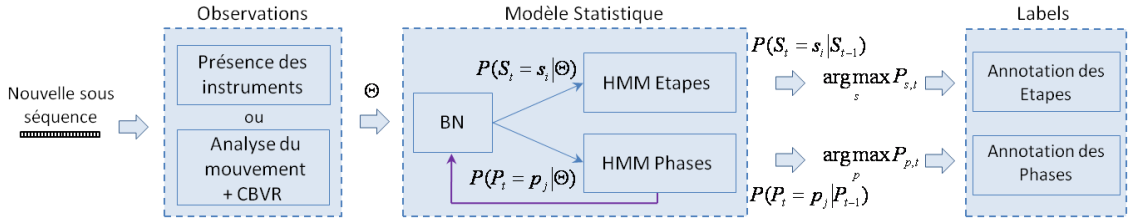


Figure 57. Méthodologie du système d'annotation automatique, en retournant les valeurs issues du HMM « phases »

IV.3.2 Caractérisation de la vidéo

Pour être annotée automatiquement la vidéo requête est découpée en sous-séquences de taille fixe qui peuvent se chevaucher. Chaque sous-séquence est alors caractérisée par son contenu visuel qui va permettre de générer les observations pour le modèle statistique. Deux approches ont été développées. La première consiste à extraire automatiquement le contenu visuel des sous-séquences (analyse du mouvement) puis à effectuer une recherche des sous-séquences les plus proches de la base de données. Le résultat de la recherche des plus proches voisins, c'est-à-dire la probabilité de réalisation de chaque label du niveau de granularité le plus fin, sera utilisé en entrée du modèle statistique. La seconde approche consiste à utiliser les informations de présence des instruments dans le champ de vue de la caméra pour générer les observations utilisées en entrée du modèle. Le modèle statistique, à partir de ces observations, va permettre de déduire les probabilités de réalisation des labels à chaque niveau de description.

IV.3.2.1 Structure de l'analyse

Nous avons choisi de découper la vidéo requête Q en un ensemble $Q = \{v_1, \dots, v_T\}$ de T sous-séquences de taille fixe $v_t = \{I_{t\,T_{\text{shift}}}, \dots, I_{t\,T_{\text{shift}}+T_{\text{scale}}}\}_{0 \le t < T}$. Comme cela est présenté dans la Figure 58. Chaque sous-séquence est enregistrée pendant T_{scale} images et l'espacement entre deux sous-séquences est notée T_{shift}. Les valeurs optimales de T_{shift} et T_{scale} sont déterminées par apprentissage (paragraphe IV.3.3.1).

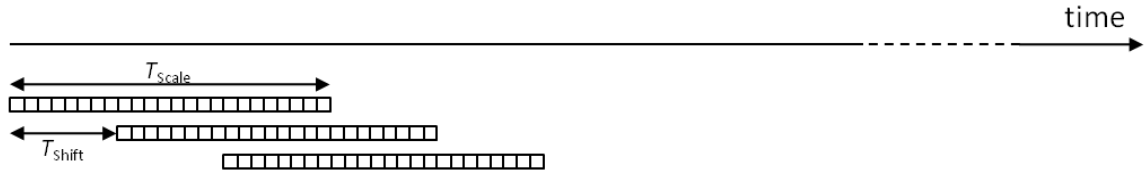


Figure 58. Découpage de la vidéo requête en en sous-séquences de taille fixe

Chacune de ces sous-séquences sera labellisée pour chaque niveau de description via l'inférence du modèle statistique. Le label sera alors assigné à chacune des T_{shift} premières images de la sous-séquence $\{I_{t\,T_{\text{shift}}}, \dots, I_{(t+1)T_{\text{shift}}}\}_{0 \le t < T}$ afin que chaque image de la vidéo soit associée à un label.

IV.3.2.2 Génération des observations

Chaque sous-séquence v_i est caractérisée selon son contenu visuel. Celui-ci va permettre de générer les observations qui fourniront les preuves au réseau bayésien, à partir desquelles vont

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

être calculées les probabilités conditionnelles. Comme le montre la Figure 59, des nœuds d'observation sont ajoutés au réseau bayésien. A chaque nouvelle sous-séquence, des nouvelles preuves sont apportées au réseau bayésien via les nœuds d'observation. Deux sources d'observations ont été testées pour évaluer notre modèle : la présence des instruments dans le champ de vue de la caméra et l'information issue de l'analyse du mouvement dans la vidéo.

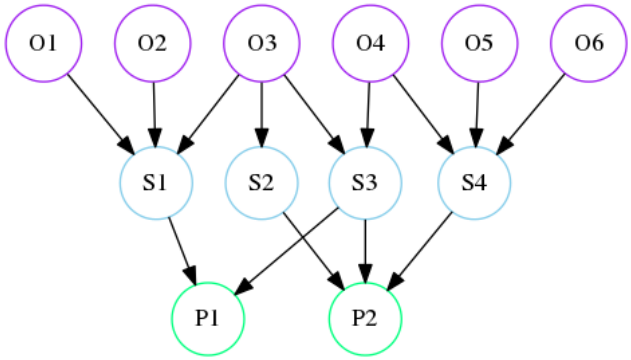


Figure 59. Exemple de réseau bayésien à deux niveaux de description, en violet les nœuds d'observation

IV.3.2.2.1 Présence des instruments

Afin de valider le modèle et d'étudier l'intérêt de la mise en place de méthodes de détection et reconnaissance des instruments chirurgicaux, nous avons utilisé la présence des instruments chirurgicaux dans le champ de vue de la caméra comme source d'observations. Nous pouvons effectivement supposer que cette information est fortement corrélée à l'exécution des étapes et des phases chirurgicales. Les images des différents instruments qu'il est possible de rencontrer dans nos vidéos de chirurgie de la cataracte sont présentées dans l'Annexe 3. Chaque nœud d'observation du réseau bayésien représente alors un instrument chirurgical. Pour connaître avec exactitude quel instrument est en cours d'utilisation, nous pouvons imaginer utiliser des radios-étiquettes (puces RFID). Cependant, il est extrêmement difficile d'intégrer ces radios-étiquettes aux instruments chirurgicaux. Pour des raisons d'asepsie, tout nouvel élément qui entre dans le bloc opératoire doit répondre en effet à des critères très stricts de stérilisation, particulièrement lorsque cela concerne les instruments chirurgicaux. Une autre approche consiste à détecter et reconnaître automatiquement des instruments chirurgicaux. Cette approche semble plus facilement réalisable, mais devra respecter les contraintes de temps réel. Dans notre travail, nous supposons ce problème résolu. Dans l'attente d'outils automatiques, nous utiliserons les informations de présence des instruments, issues d'une segmentation manuelle, présents dans la base de données. Les relations qui existent entre la présence des instruments et les étapes de la chirurgie, ainsi que les probabilités conditionnelles correspondantes, sont apprises à partir de la base de données.

IV.3.2.2.2 Analyse du mouvement

Nous nous sommes également appuyés sur des méthodes d'analyse du mouvement pour générer des observations. Pour cela nous avons utilisé les méthodes d'extraction de caractéristiques visuelles présentées dans le paragraphe III.2.1.1. Ainsi le modèle a été évalué avec deux types de signatures visuelles. Le premier type de signatures est la construction d'**histogrammes de mouvement,** calculés à partir du flux optique extrait entre deux images consécutives. Le vecteur de caractéristiques est obtenu en considérant l'ensemble des vecteurs de mouvement contenus dans la sous-séquence. Le second type de signatures constitue notre méthode de référence. Il s'agit de la construction d'**histogrammes de mots visuels**, largement utilisés dans littérature. Ces histogrammes de mots visuels sont calculés à partir des descripteurs STIP extraits tout au long de la sous-séquence (paragraphe III.2.1.1.1). La méthode a également été évaluée avec la **normalisation** les vidéos pour l'utilisation des histogrammes de mouvement comme signatures visuelles. Les vidéos ont été normalisées avec les trois types de normalisation (recalage, mise à l'échelle et sélection d'une ROI).

Une méthode de **CBVR** a ensuite été mise en place pour fournir les observations au réseau bayésien. Pour cela, les sous-séquences les plus proches de la sous-séquence requête sont recherchées dans la base de données, en comparant leurs signatures visuelles. Pour cette étape de recherche de cas les plus proches, la méthode ANN (« Approximate Nearest Neighbor Searching » en anglais) a été utilisée via la librairie[7] du même nom. Dans cette méthode, un ensemble de points (les signatures visuelles dans notre cas), appartenant à un espace de dimension D, est structuré au sein d'une structure de données qui permet de rechercher efficacement les K plus proches voisins de tout point requête q. La méthode ANN suppose que les distances sont mesurées avec une distance de Minkowski (distance de Manhattan, distance euclidienne, etc…). Nous utilisons ici une distance euclidienne. Le nombre K optimal de plus proches voisins est déterminé par apprentissage (paragraphe IV.3.3.1). Le résultat de la recherche des plus proches voisins fournit la probabilité de réalisation des différentes étapes chirurgicales. Les probabilités de réalisations ainsi obtenues vont fournir les preuves au réseau bayésien. Cependant, le réseau bayésien fonctionne avec des preuves binaires (vrai ou faux) et non avec des preuves probabilistes. C'est pourquoi chaque observation est associée à un intervalle de probabilités (par exemple : la recherche des K plus proches voisins indique que la probabilité de réalisation de l'étape « Incision » est comprise entre 10% et 20% pour la sous-séquence requête). Chaque nœud d'observation représente alors un intervalle de probabilité de réalisation d'une étape chirurgicale donnée. Il sera alors nécessaire de déterminer le nombre optimal $nbObs$ d'intervalles.

La construction et l'utilisation du modèle statistique permettant de déduire les probabilités de réalisation pour chacun des niveaux de description à partir des observations ainsi calculées sont présentées dans le paragraphe suivant (IV.3.1).

IV.3.3 Evaluation

Les différentes versions du modèle ont été évaluées via une validation croisée. En effet, les modèles statistiques utilisés nécessitent une base d'apprentissage contenant un nombre de cas suffisant pour permettre de modéliser l'ensemble des relations qui peuvent exister. Or notre base de données, interprétée à plusieurs niveaux de description, est constituée de 30 vidéos. Ainsi, comme cela est présenté dans la Figure 60, la base de données a été découpée en 6 sous-bases de tailles égales. Le modèle a alors été évalué pour les 6 sous-bases, en utilisant à chaque fois les 5 sous-bases restantes pour l'apprentissage.

[7] http://www.cs.umd.edu/~mount/ANN/

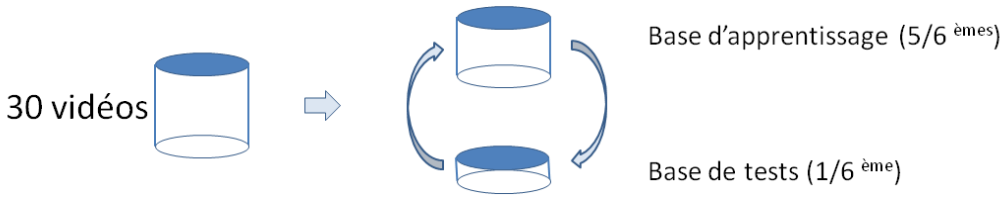


Figure 60. Utilisation d'une validation croisée pour l'évaluation du modèle

La base d'apprentissage est utilisée pour apprendre les différents blocs du modèle statistique (réseau bayésien, HMM, CRF), mais également pour fixer les valeurs optimales des différents paramètres du modèle (T_{shift}, T_{scale}, etc…).

IV.3.3.1 Optimisation des paramètres

Il est nécessaire de réaliser un apprentissage pour déterminer les paramètres optimaux qui permettent d'obtenir la meilleure reconnaissance possible. Ces paramètres sont notamment les valeurs T_{shift} et T_{scale} qui déterminent la taille des sous-séquences et l'écart temporel entre ces sous-séquences. Dans le cas de l'analyse du mouvement comme source d'observations pour le modèle, il est également nécessaire de déterminer le nombre idéal de K plus proches voisins ainsi que le nombre $nbObs$ de classes nécessaires pour créer les nœuds d'observation. la base d'apprentissage a elle-même été divisée en deux sous-bases de tailles quasi-égales (12 et 13 vidéos). La recherche des paramètres optimaux a été effectuée selon une méthode de recherche par quadrillage (*grid search*). A chaque itération de la méthode, on fait varier un à un les paramètres pendant que les autres sont fixes. La valeur de chaque paramètre permettant d'obtenir les meilleures performances de reconnaissance, évaluées en mesurant l'aire moyenne sous la courbe ROC (paragraphe III.4.2) est retenue. Lorsque les valeurs optimales n'évoluent plus d'une itération à l'autre, les valeurs sont retenues définitivement. Un jeu de paramètres optimal a ainsi été déterminé pour chacune des six validations. Cette procédure d'apprentissage, qui fait intervenir deux validations croisées imbriquées, est complexe mais garantit des résultats non biaisés.

IV.3.3.1.1 Utilisation des HMM

Dans le cas de l'utilisation des modèles de Markov pour modéliser le déroulement temporel de la chirurgie, il est nécessaire de déterminer les pas de temps optimaux auxquels on détermine les probabilités de transition : hmm_{S} pour les labels « étapes » et hmm_{P} pour les labels « phases ».

Une première validation du modèle a été effectuée en utilisant comme observations la présence des instruments chirurgicaux dans le champ de vue de la caméra. Cette source d'observations étant la plus fiable, cette configuration a été utilisée pour déterminer les valeurs optimales pour les paramètres T_{shift}, T_{scale}, hmm_{S} et hmm_{P}. Les paramètres optimaux obtenus sont présentés dans le Tableau 12. Ces valeurs seront conservées pour l'utilisation de l'analyse du mouvement comme source d'observations.

Tableau 12. Paramètres optimaux obtenus pour l'utilisation de la présence des instruments comme observations

sous-base	1	2	3	4	5	6
T_{scale}	50	50	50	50	50	50
T_{shift}	25	25	25	25	25	25
hmm_{S}	10	7	9	8	6	5
hmm_{P}	10	55	10	50	30	25

On observe une grande variabilité du paramètre hmm_{P}. Il se peut que ce paramètre influe peu le résultat, et que, par conséquent la sélection du paramètre optimal dépend du hasard. Les paramètres optimaux propres à l'utilisation de l'analyse du mouvement comme source d'observations pour le modèle statistique ont ensuite été déterminés. Il a été nécessaire de fixer le nombre optimal de K plus proches voisins, et le nombre de classes nécessaires pour la création des nœuds d'observation ($nbObs$). Ces valeurs ont été évaluées pour l'utilisation des histogrammes de mouvements (HM) comme signatures visuelles, puis pour l'utilisation des histogrammes de mots visuels (BoW). Les valeurs optimales obtenues sont présentées dans le Tableau 13.

Tableau 13. Paramètres optimaux obtenus pour l'utilisation de l'analyse du mouvement dans la vidéo comme source d'observations

sous-base	1	2	3	4	5	6
			HM			
nbObs	5	5	3	3	3	3
K	150	150	150	150	150	150
			BoW			
nbObs	4	4	4	4	4	4
K	49	41	49	51	51	53

Dans le cas de l'évaluation du modèle avec le retour du HMM « phases » vers le réseau bayésien, il est nécessaire de déterminer le nombre de classes optimal pour la création des nœuds d'observation. Ce nombre a alors été réévalué. Cette évolution du modèle a été évaluée pour l'utilisation de la présence des instruments comme observations et l'analyse du mouvement via la construction d'histogrammes de mots visuels. Les valeurs optimales obtenues sont présentées dans le Tableau 14.

Tableau 14. Paramètres optimaux obtenus pour le retour des résultats de l'inférence du HMM « phases » dans le réseau bayésien

sous-base	1	2	3	4	5	6
			HM			
nbObs	4	4	5	3	6	5
			Instruments			
nbObs	4	6	7	6	4	3

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

IV.3.3.1.2 Utilisation des CRF

L'utilisation des CRF à la place des HMM a également introduit de nouveaux paramètres. Il a notamment été nécessaire de déterminer les valeurs L et Δ_s qui détermine l'ensemble de sous-séquences présentes et passées $\{v_{i-l\Delta_s}\}_{0\leq l\Delta_s\leq L}$ utilisé pour la construction des potentiels. Il été nécessaire de regrouper les observations en un nombre de classes fini ($nbClasses$). La méthode a été évaluée avec les deux types d'observations (présence des instruments chirurgicaux et analyse du mouvement). Dans le cas de l'analyse du mouvement, seules les signatures visuelles sous forme d'histogrammes de mouvement ont été utilisées. Les valeurs des paramètres T_{shift}, T_{scale}, $nbObs$ et K utilisées sont identiques aux valeurs choisies dans le cas de l'utilisation des HMM. Les valeurs optimales des paramètres $nbClasses$, L et Δ_s ont été évaluées, pour chaque validation, pour chaque niveau de description (phases et étapes).

Tableau 15. Paramètres optimaux obtenus pour l'utilisation de l'analyse du mouvement comme source d'observations avec des histogrammes de mouvement comme signatures visuelles

	sous-base	1	2	3	4	5	6
Phases	nbClasses	4	18	2	4	8	8
	L	1	85	40	35	5	1
	Δ_s	1	58	26	12	2	1
Etapes	nbClasses	1	19	20	15	18	6
	L	15	50	55	65	60	35
	Δ_s	1	48	32	28	4	1

Les paramètres optimaux obtenus dans le cadre de l'utilisation de l'analyse du mouvement comme source d'observations (comparaison d'histogrammes de mouvement) Tableau 15. Les paramètres optimaux obtenus dans le cadre de l'utilisation de la présence des instruments dans le champ de vu de la caméra comme source d'observations sont présentés dans le Tableau 16.

Tableau 16. Paramètres optimaux obtenus dans le cadre de l'utilisation de l'information de présence des instruments comme source d'observations

	sous-base	1	2	3	4	5	6
Phases	nbClasses	12	4	2	2	2	17
	L	50	1	30	5	1	70
	Δ_s	38	1	1	4	1	50
Etapes	nbClasses	10	20	18	16	16	18
	L	10	40	40	5	60	35
	Δ_s	1	2	18	1	34	2

Ces paramètres optimaux, obtenus par validation croisées sur les 6 bases d'apprentissage, vont être utilisés pour les différentes évaluations de la méthode sur les 6 bases de test.

IV.3.3.2 Inférence

Lors de l'analyse d'une vidéo requête, l'inférence du modèle permet de déterminer les labels les plus probables à chacun des niveaux de description, pour chaque sous-séquence enregistrée. Le modèle prend en entrée les observations enregistrées pour la nouvelle sous-séquence. Le réseau bayésien est alors parcouru en fonction des nouvelles preuves obtenues et fournit les probabilités de réalisation de chacun des labels pour la sous-séquence. Ces probabilités sont alors interprétées par deux HMM à pas de temps fixes (respectivement hmm_{S} et hmm_{P}) ou par deux CRF.

IV.3.3.2.1 Réseau bayésien seul

L'inférence du réseau bayésien consiste à calculer la probabilité d'une variable étant donné un ensemble d'observations. Il existe pour cela différentes méthodes exactes ou approchées. La méthode historique, appelée « message passing », proposée par J. Pearl [83] est une méthode d'inférence exacte. Elle consiste en une propagation par messages de nœud en nœud dans un arbre. Ainsi, chaque nœud envoie des messages à ces voisins. Cette méthode a été généralisée aux graphes (plus généraux) via la construction d'un arbre joint (algorithme « Junction tree ») [84]. Cette méthode d'inférence est largement utilisée pour l'inférence des réseaux bayésiens. Cependant dans le cas de réseaux bayésiens complexes, l'inférence du réseau peut être coûteuse en temps de calcul. Au vu de la complexité de la structure de notre réseau bayésien, et de nos objectifs d'analyse en temps réel de la chirurgie, nous nous sommes orientés vers l'utilisation d'une méthode approchée. Nous nous sommes appuyés sur une méthode stochastique via l'échantillonneur de Gibbs [85]. Cette méthode appartient à la famille des méthodes d'inférence MCMC (Markov chain Monte Carlo). Elle revient à choisir à chaque itération une variable dont on change la valeur en fonction des valeurs de son entourage [80]. Soit le réseau bayésien à N variables $\mathcal{V} = \{S_1, \dots, S_N\}$ dont certaines sont observées, l'algorithme fonctionne alors de la manière suivante :

- *Initialisation* : pour toute variable on choisit une valeur (compatible avec les observations) aléatoirement

- *A chaque itération* k : on cherche à calculer S_k en fonction de l'échantillon précédent S_{k-1} Pour cela, on choisit une variable S_i (non observées) et on modifie sa valeur en fonction de sa loi conditionnellement à ses parents dans le graphe.

La librairie DLib[8] a été utilisée pour réaliser l'inférence de réseau bayésien via l'échantillonneur de Gibbs.

[8] http://dlib.net/

IV.3.3.2.2 HMM

L'inférence des modèles de Markov se fait via l'algorithme de Viterbi [78] qui cherche la séquence d'états cachés (labels) la plus probable. Cela revient à chercher le chemin le plus probable dans le graphe, c'est-à-dire, pour une séquence d'observations $O = (o_1, \dots, o_T)$ la probabilité $P(S_1, \dots, S_T|O, \lambda)$ maximale, avec $\lambda = \{A, B, \pi\}$ les paramètres du modèle. On définie alors :

$$\delta_t(s_i) = \max_{S_1,\dots,S_{t-1}} P(S_1, \dots, S_{t-1}, o_1, \dots, o_T \mid \lambda)$$

la probabilité maximale des chemins se terminant par le label $s_i \in \mathcal{S} = \{(s_k)_{0\leq k<N_s}\}$. L'algorithme fonctionne alors de la manière suivante :

- *Initialisation* : $\delta_1(s_i) = \pi_i b_i(o_1)$
- A chaque pas de temps t : $\delta_t(s_j) = \max_{s_i \in \mathcal{S}} [\delta_{t-1}(s_i)a_{ij}]\, b_j(o_t)$
- Evaluation du chemin le plus probable : $P^* = \max_{s_i \in \mathcal{S}} [\delta_{t-1}(s_i)]$

Les probabilités $b_j(o_t) = P(s_j|o_t)$ sont fournies par le réseau bayésien. Le fonctionnement est le même pour déterminer la séquence de label « phases » la plus probable.

IV.3.3.2.3 CRF

L'inférence des CRF se fait généralement via un algorithme « forward-backward »[86]. Cependant nous souhaitons réaliser une analyse « en direct », et nous ne pouvons donc nous appuyer que sur des informations présentes ou passées. Nous avons alors utilisé un algorithme « forward » (méthode dite « avant ») pour déterminer la probabilité $P(S_t = s_i)$ d'obtenir le label $s_i \in \mathcal{S} = \{(s_k)_{0\leq k<N_s}\}$ pour la sous-séquence enregistrée au pas de temps t [51]:

$$\begin{cases} \log\big(P(S_0 = s_i)\big) = \sum_{u=1}^{U} \lambda_u \psi_0^u(s_i, \boldsymbol{o}_t) \\ \log\big(P(S_t = s_i)\big) = \max_{s_j \in \mathcal{S}} \left[\log\Big(P\big(S_{t-1} = s_j\big)\Big) + \sum_{u=1}^{U} \lambda_u \psi_t^u(s_i, \boldsymbol{o}_t) + \sum_{b=1}^{N_s} \mu_u \psi_{t-1,t}^b\big(s_j, s_i, \boldsymbol{o}_t\big) \right] \end{cases}$$

Il en a été de même pour déterminer la probabilité $P(P_t = p_i)$ d'obtenir le label $p_i \in \mathcal{P} = \{(p_j)_{0\leq j<N_p}\}$. La librairie Wapiti[9] [82] a été utilisée pour réaliser l'inférence des CRF. Cette dernière a été modifiée pour permettre l'inférence « en direct » via l'algorithme « forward » [51].

[9] https://wapiti.limsi.fr/

IV.3.3.3 Résultats

Différentes versions du modèle ont été évaluées. La version initiale est composée d'un réseau bayésien et de modèles de Markov cachés. Une seconde version a été évaluée pour laquelle le résultat de l'inférence du réseau bayésien au niveau « phases » a été retournée vers le réseau bayésien. Nous avons également comparé l'utilisation des HMM avec l'utilisation de CRF qui semble donner de meilleurs résultats dans le cadre de l'analyse de vidéos chirurgicales [30][51]. Les différentes évaluations ont été réalisées avec deux types d'observations : la présence des instruments chirurgicaux dans le champ de vue de la caméra et l'analyse du mouvement associé à une recherche de plus proches voisins.

Les performances de reconnaissance pour les labels des deux niveaux de description utilisés on été mesurées en termes d'aire sous la courbe ROC (paragraphe III.4.2 page 85). Pour chaque pli de la validation croisée, une courbe ROC a été construite pour chaque label. Nous présentons dans chaque tableau, pour les deux niveaux (étapes et phases), la moyenne des aires obtenues sous l'ensemble des courbes ROC construites pour les six bases de test. La moyenne des résultats obtenus pour les deux niveaux est également présentée (A_z Moyenne). Enfin, le nombre d'images qu'il est possible de traiter par seconde avec le système, obtenu avec un cœur d'un processeur « quad core » Intel(R) Core(TM) i7-3770 (3.40GHz), est également présenté dans la dernière ligne des tableaux.

IV.3.3.3.1 Réseau Bayésien

Dans un premier temps nous avons évalué les méthodes d'élagage du réseau bayésien présentées dans le paragraphe IV.3.1.1. En effet, plus le réseau bayésien est complexe, plus son inférence est coûteuse en temps de calcul. De plus le nombre de combinaisons possibles des nœuds parents pour le calcul des probabilités conditionnelles de chaque nœud augmente de façon exponentielle par rapport au nombre de parents, ce qui nécessite l'utilisation d'une plus grande base d'apprentissage. Nous avons, dans cette optique, cherché à réduire le nombre d'arcs dans le graphe, en limitant le nombre de parents possibles pour chaque nœud. Pour cela, nous avons évalué les deux critères de sélection des arcs pertinents présentés dans le paragraphe IV.3.1.1, afin de sélectionner les 10 nœuds parents les plus pertinents. Le premier critère consiste à sélectionner les 10 arcs pour lesquels l'information mutuelle (MI) est maximale. Ce choix est imposé par les contraintes de temps de calcul (pour l'inférence) et permet de limiter la taille des tables de probabilité associées au nœud à une dimension raisonnable. En effet, avant élagage, certains nœuds peuvent atteindre une vingtaine de parents. Le second critère consiste à sélectionner les 10 arcs pour lesquels le nombre N_{ij1} est maximal. Le nombre N_{ij1} représente le nombre d'instances dans la base où le nœud S_i est réalisé et dont le nœud parent $\mathrm{Pa}(S_i) = \{S_j\}$ est également réalisé. Nous avons évalué ces critères avec les résultats de la recherche des cas les plus proches comme source d'observations (via la comparaison d'histogrammes de mouvement). En effet, le graphe est plus complexe dans cette configuration que pour l'utilisation de la présence des instruments comme source d'observations. L'utilisation d'une étape d'élagage de l'arbre est donc plus pertinente dans ce cas. Les résultats, en termes d'aire moyenne sous la courbe ROC, sont présentés dans le Tableau 17.

Tableau 17. Comparaison des méthodes d'élagage du réseau bayésien

	10 MI max	10 N_{ij1}max
A_z Etapes	0,637	0,637
A_z Phases	**0,603**	0,611
A_z Moyenne	**0,620**	0,624
Nb image /s	12,4	13,0

Les résultats obtenus sont faibles, car nous n'utilisons pas dans ce cas d'informations sur le déroulement temporel. Néanmoins, les meilleurs résultats ont été obtenus en choisissant comme critère le nombre N_{ij1}. Ce critère sera donc utilisé pour la construction du réseau bayésien pour les évaluations suivantes. Nous pourrions néanmoins réfléchir à une utilisation conjointe des deux critères. Nous pourrions également réfléchir à l'utilisation d'une méthode d'apprentissage des paramètres du graphe permettant de gérer des données incomplètes, afin d'avoir une estimation plus fiable des probabilités conditionnelles.

IV.3.3.3.2 Réseau Bayésien et HMM

Nous avons évalué une première version du modèle, pour laquelle un réseau bayésien a été combiné avec des modèles de Markov. Afin d'évaluer les différents blocs du modèle, celui-ci a été évalué pour différentes sources d'observations, en comparant :

- l'utilisation du réseau bayésien seul (BN)
- l'utilisation du réseau bayésien combiné avec un HMM pour le niveau de description « étapes » (BN + HMM S)
- l'utilisation du réseau bayésien combiné avec un HMM pour le niveau de description « phases » (BN + HMM P)
- l'utilisation du réseau bayésien combiné avec les deux HMM (BN + HMM S&P)

Le modèle a été évalué dans un premier temps avec comme source d'observations la **présence des instruments** dans le champ de vue de la caméra. Cette source d'observations est la plus fiable, elle permet donc de valider les différents éléments du modèle. Les performances en termes d'aire moyenne sous la courbe ROC sont présentées dans le Tableau 18.

Tableau 18. Evaluation du modèle combinant un réseau bayésien et des HMM, avec comme source d'observations la présence des instruments dans le champ de vue de la caméra

	BN seul	BN + HMM S	BN + HMM P	BN HMM S&P
A_z Etapes	**0,931**	0,906	0,930	0,903
A_z Phases	0,813	0,817	0,919	**0,922**
A_z Moyenne	0,874	0,861	**0,925**	0,913
Nb image /s	21,7	21,8	21,4	21,5

Nous pouvons constater que le réseau bayésien permet de déduire le label le plus probable pour les étapes et les phases avec de bonnes performances. L'aire moyenne obtenue pour l'ensemble des labels (étapes et phases) est de $0{,}874$. L'utilisation d'un HMM pour le niveau de description en phases chirurgicales améliore les performances de reconnaissances. En revanche, les performances ne sont pas améliorées par l'utilisation d'un HMM pour le niveau de description en étapes. Cela peut s'expliquer par le fait que les relations entre les différentes étapes sont complexes et que notre base d'apprentissage est de petite taille. Il est alors possible que certaines transitions ne soient pas représentées dans la base d'apprentissage. Cela devrait être résolu par l'utilisation de CRF à la place des HMM. Nous constatons également que l'amélioration des performances de reconnaissance au niveau des phases, par l'utilisation d'un HMM à ce niveau, ne permet pas d'améliorer les performances de reconnaissance des labels « étapes ». Cela s'explique par le fait que les informations issues de l'inférence des HMM ne sont pas retransmises au réseau bayésien. L'information obtenue ne peut alors pas remonter vers le niveau de description en étapes chirurgicales. Cela devrait être résolu par la mise en place d'un retour de l'information issue de l'inférence du HMM vers le réseau bayésien (paragraphe IV.3.3.3.3). **Enfin, le système permet de traiter environ 21 images par seconde et est donc compatible avec une utilisation en temps réel.**

Le modèle a ensuite été évalué en utilisant l'**analyse du mouvement** comme source d'observations. Dans un premier temps, l'utilisation de signatures visuelles sous forme d'*histogrammes de mouvement* a été évaluée. Les observations sont générées par le résultat de la recherche des sous-séquences les plus proches dans la base d'apprentissage. Les performances, en termes d'aire moyenne sous la courbe ROC, sont présentées dans le Tableau 19.

Tableau 19. Evaluation du modèle combinant un réseau bayésien et des HMM, avec comme source d'observations les résultats de la recherche des cas les plus proches (comparaison d'histogrammes de mouvement)

	BN seul	BN + HMM S	BN + HMM P	BN HMM S&P
A_z **Etapes**	0,637	**0,677**	0,638	0,674
A_z **Phases**	0,611	0,608	**0,814**	0,812
A_z **Moyenne**	0,624	0,643	0,726	**0,743**
Nb image /s	13,0	13,6	13,6	13,2

Nous pouvons constater que les performances obtenues par l'inférence du réseau bayésien sont faibles (avec une aire moyenne sous la courbe ROC de $0{,}624$). Cela est lié au fait que les observations, issues de la recherche des K plus proches sous-séquences dans la base d'apprentissage, sont moins fiables que la présence des instruments (fournie manuellement par le chirurgien). Ces performances sont améliorées par l'utilisation des HMM aux deux niveaux de description, ce qui permet d'obtenir une aire moyenne sous la courbe ROC de $0{,}743$ et les performances de reconnaissance des labels « phases » sont satisfaisantes avec une aire moyenne sous la courbe ROC de $0{,}812$. Les temps de calcul sont plus longs que dans le cas de l'utilisation de la présence des instruments comme source d'observations. Cela s'explique par le fait que des étapes supplémentaires sont ajoutées au système : l'extraction des caractéristiques visuelles et la recherche des K plus proches voisins. Néanmoins, le système reste très rapide et permet d'analyser environ 13 images par seconde.

Afin d'améliorer la pertinence des observations fournies en entrée du modèle, nous avons cherché à améliorer les performances de la recherche des K plus proches voisins. Pour cela, nous avons utilisé la **normalisation des vidéos**, évaluée dans les paragraphes III.2.1.2 et III.4.3.2. Les vidéos ont été normalisées en recalant spatialement les images de la vidéo, en les mettant à la même échelle et en sélectionnant une région d'intérêt (ROI). La comparaison des résultats avec et sans normalisation des vidéos est présentée dans le Tableau 20. Pour cette comparaison, les signatures visuelles utilisées sont les histogrammes de mouvement.

Tableau 20. Evaluation du modèle combinant un réseau bayésien et des HMM, avec comme source d'observations les résultats de la recherche des cas les plus proches (comparaison d'histogrammes de mouvement) ; Comparaison avec et sans normalisation des vidéo (recalage spatial, sélection d'une ROI et mise à l'échelle)

	Sans normalisation		Avec normalisation	
	BN + HMM P	**BN + HMM S&P**	**BN + HMM P**	**BN + HMM S&P**
A_z Etapes	0,638	0,674	0,693	**0,721**
A_z Phases	0,814	0,812	**0,836**	0,832
A_z Moyenne	0,726	0,743	0,764	**0,777**
Nb image /s	**13,6**	**13,2**	11,31	11,22

Nous pouvons constater que, comme nous l'espérions, les performances ont été améliorées par la normalisation des vidéos, bien que cela augmente légèrement les temps de calcul (environ 11 images traitées par seconde au lieu de 13).

Le modèle a ensuite été évalué avec l'utilisation d'*histogrammes de mots visuels* comme signatures des sous-séquences. Les histogrammes de mots visuels sont largement utilisés dans la littérature et ils ont permis d'obtenir de meilleures performances de reconnaissance dans le cadre de l'évaluation de la mesure de similitude de Piciarelli et al. [44] que nous avons adaptée à la comparaison de vidéos médicales (paragraphe III.4.3.2). Les performances de reconnaissances ont été évaluées de façon similaire à l'évaluation des sources d'observations précédemment utilisées. Les résultats, en termes d'aire moyenne sous la courbe ROC, sont présentés dans le Tableau 21.

Tableau 21. Evaluation du modèle combinant un réseau bayésien et des HMM, avec comme source d'observations les résultats de la recherche des cas les plus proches (comparaison d'histogrammes de mots visuels)

	BN seul	**BN + HMM S**	**BN + HMM P**	**BN HMM S&P**
A_z Etapes	0,636	0,677	0,638	**0,686**
A_z Phases	0,627	0,627	**0,834**	0,828
A_z Moyenne	0,632	0,652	0,736	**0,757**
Nb image /s	0,92	0,92	0,92	0,92

Nous pouvons constater que les résultats sont légèrement améliorés par rapport l'utilisation des histogrammes de mouvement. Cependant, les temps de calcul sont beaucoup plus importants

et le système traite moins d'une image par seconde, ce qui n'est pas compatible avec une utilisation temps réel du système. Cela est dû à l'extraction des points d'intérêts STIP qui est coûteuse en temps de calcul (paragraphe III.4.3.3).

En conclusion, les meilleurs résultats ont été obtenus avec l'utilisation de la présence des instruments chirurgicaux dans le champ de vu de la caméra. Cela était attendu car cette source d'observations est fiable (car fournie manuellement par les médecins) et fortement corrélée avec la réalisation des étapes et des phases chirurgicales. Ces résultats sont encourageants. Ils ont permis de valider le modèle et de montrer que l'on était capable de remonter d'une information bas niveau vers les descriptions en étapes et en phases chirurgicales. Il semble alors pertinent de mettre en place, par la suite, des méthodes de reconnaissance automatique des instruments chirurgicaux. Cette évaluation a également permis de montrer certaines limites des HMM dans notre situation. En effet, la faible taille de la base d'apprentissage ne permet pas de représenter toutes les transitions possibles. Cet aspect devrait être résolu par l'utilisation des CRF. L'utilisation de l'analyse du mouvement via des histogrammes de mouvement comme source d'observation a permis d'obtenir des résultats satisfaisants avec l'utilisation de vidéos normalisées. Enfin l'utilisation des signatures sous forme d'histogrammes de mots visuels permet d'obtenir de meilleures performances de reconnaissance. Cependant, dans ce cas, le système n'est pas compatible avec une utilisation en temps réel. Il pourrait cependant s'envisager pour l'indexation automatique d'archives de vidéos enregistrées durant les interventions. Pour la suite de nos évaluations, nous nous appuierons donc sur la présence des instruments et sur la construction d'histogrammes de mouvement (associée à une recherche des cas les plus proches).

IV.3.3.3.3 Réseau Bayésien et retour HMM « phases »

Nous avons constaté dans le paragraphe précédent que l'amélioration des performances de reconnaissance au niveau des phases par l'utilisation d'un HMM n'est pas retransmise au réseau bayésien. Or cette information pourrait permettre d'améliorer les performances de reconnaissance des labels « étapes ». En effet, la connaissance que nous avons sur les labels « phases » les plus probables peut nous aider à déterminer les labels « étapes » (déduction des causes à partir des conséquences). Pour cela nous avons mis en place un retour de cette information vers le réseau bayésien (paragraphe IV.3.1.2).

Dans un premier temps, cette évolution du modèle a été évaluée avec la présence des instruments qui est la source d'observation la plus fiable. La comparaison du modèle avec et sans retour des informations du HMM au niveau des phases sont présentés dans le Tableau 22.

Tableau 22. Comparaison du modèle combinant un réseau bayésien et des HMM, avec retour du résultat de l'inférence du HMM pour le niveau « phases » vers le réseau bayésien ; utilisation de la présence des instruments comme source d'observations

	Sans retour		Avec retour	
	BN + HMM P	**BN + HMM S&P**	**BN + HMM P**	**BN + HMM S&P**
A_z Etapes	0,930	0,903	0,961	**0,946**
A_z Phases	0,919	**0,922**	0,910	0,910
A_z Moyenne	0,925	0,913	**0,935**	0,928
Nb image /s	**21,4**	**21,5**	16,4	16,9

Katia CHARRIERE
Laboratoire de Traitement de l'Information Médicale (LaTIM – UMR 1101)
Télécom Bretagne – Département ITI

Nous pouvons constater que dans cette configuration les performances de reconnaissances au niveau des étapes sont améliorées par les observations obtenues au niveau des phases (résultats du HMM). Le modèle est donc capable de remonter des informations des conséquences vers les causes. En revanche, dans cette configuration, le système ne peut traiter en moyenne que 16 images par seconde au lieu de 21 sur notre ordinateur. Cela s'explique par la plus grande complexité du réseau bayésien.

Nous avons également étudié le retour du HMM au niveau de description en phases vers le réseau bayésien en utilisant comme source d'observation l'analyse du mouvement dans la vidéo via des histogrammes de mouvement calculés sur des vidéos normalisées. Les résultats de la comparaison des deux systèmes (avec et sans retour) sont présentés dans le Tableau 23.

Tableau 23. Comparaison du modèle combinant un réseau bayésien et des HMM, avec retour du résultat de l'inférence du HMM pour le niveau de description en phases vers le réseau bayésien ; utilisation des résultats de la recherche des cas les plus proches comme source d'observations (comparaison d'histogrammes de mouvement)

	HMM P	HMM P + norm.	HMM P + norm. + retour HMM	HMM S&P	HMM S&P + norm.	HMM S&P + norm. + retour HMM
A_z Etapes	0,638	0,693	0,705	0,674	0,721	**0,733**
A_z Phases	0,814	**0,836**	0,818	0,812	0,832	0,819
A_z Moyenne	0,726	0,764	0,762	0,743	**0,777**	0,776
Nb image /s	**13,6**	11,3	9,8	**13,2**	11,2	9,8

Nous pouvons constater que dans cette situation le retour du HMM vers le réseau bayésien permet également d'améliorer les performances de reconnaissance mais de façon plus contrastée que dans le cas de l'utilisation de la présence des instruments comme source d'observations.

IV.3.3.3.4 Réseau Bayésien et CRF

Le HMM a montré certaines limites pour le niveau de description en étapes chirurgicales car la faible taille de la base d'apprentissage ne permet pas de représenter toutes les transitions possibles. C'est pourquoi nous avons choisi d'évaluer une autre modélisation du déroulement temporel de la chirurgie, qui semble plus adaptée à l'utilisation d'une petite base de données : les CRF. Les résultats de la comparaison des deux méthodes, obtenus avec la présence des instruments et l'analyse du mouvement (via la construction d'histogrammes de mouvement) comme sources d'observations, sont présentés dans le Tableau 24.

Tableau 24. Comparaison du modèle combinant un réseau bayésien et des HMM avec le modèle combinant un réseau bayésien et des CRF, avec comme source d'observations la présence des instruments, puis les résultats de la recherche des cas les plus proches (comparaison d'histogrammes de mouvement)

	Instruments	MH

	BN + HMM S&P	BN + CRF S&P	BN + HMM S&P	BN + CRF S&P
A_z Etapes	0,903	**0,980**	0,674	**0,691**
A_z Phases	0,922	**0,986**	0,812	**0,828**
A_z Moyenne	0,913	**0,983**	0,743	**0,759**
Nb image /s	21,5	21,7	13,2	13,2

Les résultats ont été améliorés par l'utilisation des CRF, particulièrement dans le cas de l'utilisation de la présence des instruments comme source d'observations. Dans ce cas, des très bons taux de reconnaissances ont été obtenus avec une aire moyenne sous la courbe ROC de $0{,}983$ au lieu de $0{,}913$. La reconnaissance est bonne aussi bien pour la description en phases que pour la description en étapes chirurgicales. Les CRF semblent donc plus adaptés que les HMM pour la modélisation temporelle du processus chirurgical. Les résultats ont été améliorés, mais de façon moins importante dans le cas de l'utilisation des signatures l'analyse du mouvement via la construction d'histogrammes de mouvement comme source d'observations, avec une aire moyenne sous la courbe ROC de $0{,}691$ au lieu de $0{,}674$.

IV.3.3.4 Conclusion

Le système de séquençage multi-échelles a été validé avec différentes configurations. De très bons résultats ont été obtenus avec présence des instruments comme source d'observations, particulièrement avec l'utilisation des CRF pour modéliser le processus temporel de la chirurgie. Dans cette configuration, nous obtenons une labellisation quasi-parfaite, avec une aire moyenne sous la courbe ROC de $0{,}982$. La reconnaissance est bonne pour les phases, mais également pour les étapes chirurgicales pour lesquelles processus chirurgical est complexe, avec un grand nombre de transitions possibles. Il est donc possible de séquencer automatiquement une chirurgie requête en étapes et en phases chirurgicales, à partir de la connaissance des instruments chirurgicaux utilisés. Il semble donc intéressant de mettre en place une méthode de reconnaissance automatique des instruments chirurgicaux. Cependant cette tâche est complexe et peut être coûteuse en temps de calcul. Des résultats plus contrastés ont été obtenus pour l'analyse du mouvement. De bons résultats ont été obtenus pour la labellisation des phases, avec notamment une aire moyenne sous la courbe ROC de $0{,}828$. En revanche, dans cette configuration, les performances de reconnaissance pour la labellisation en étapes chirurgicales sont plus faibles avec une aire moyenne sous la courbe ROC de $0{,}691$. Il serait intéressant d'évaluer le modèle avec les CRF avec des vidéos normalisées et en intégrant le retour du résultat du CRF vers le réseau bayésien. En effet, dans le cas de l'utilisation des HMM, la normalisation des vidéos et le retour des résultats du HMM pour le niveau de description en phase vers le réseau bayésien ont permis d'améliorer les performances de reconnaissance des labels « étapes » en passant d'une aire moyenne sous la courbe ROC de $0{,}674$ à une aire moyenne de $0{,}733$. Cependant, ces adaptations nécessitent un peu plus de temps de calcul (9 images traitées par seconde au lieu de 13). Enfin, le système est compatible avec une utilisation en temps réel, avec 21 images traitées par seconde pour l'utilisation de la présence des instruments chirurgicaux comme source d'observations, et 13 images par seconde pour l'utilisation de l'analyse du mouvement (via des histogrammes de mouvement).

IV.4 Discussion – Conclusion

Deux approches originales ont été évaluées pour le séquençage automatique des vidéos de chirurgie. Ces deux approches s'appuient sur une modélisation statistique du processus chirurgical qui apporte une information contextuelle lors du choix du label. La première approche utilise une modélisation en arbres de la chirurgie, alors que la seconde approche utilise des graphes pour construire un modèle multi-échelles de la chirurgie.

La première méthode évaluée est une approche basée sur la construction d'un arbre des déroulements possibles de la chirurgie. Chaque nœud de l'arbre représente une séquence moyenne de chirurgie et la structure de l'arbre représente les différents ordonnancements possibles des séquences moyennes. Cette méthode offre donc une alternative plus générale à la construction d'une chirurgie moyenne via l'algorithme DTW, en autorisant des déroulements multiples de la chirurgie. Elle permet également une structuration de l'espace de recherche. Deux méthodes de construction ont été évaluées : une méthode non supervisée et une méthode supervisée s'appuyant sur les annotations des vidéos apportées par les chirurgiens. Dans les deux cas, il n'a pas été possible d'aboutir à un séquençage automatique concluant. Lors de la comparaison des vidéos entre elles ou lors de la comparaison d'une vidéo avec une séquence moyenne, il n'y a pas eu de divergence observée. Cela est lié à la grande variabilité des signatures visuelles et à leur caractère peu reproductible. Des améliorations pourraient être apportées au modèle pour pallier ce problème. Il pourrait être intéressant par exemple de créer des nœuds qui regroupent un ensemble de séquences et non pas des nœuds représentés par une séquence moyenne. La comparaison d'une vidéo requête avec le modèle se ferait alors par une recherche des séquences les plus proches parmi les groupes de séquences (nœuds) possibles.

La seconde méthode utilise le caractère multi-échelles de notre nouvelle description de la chirurgie afin de séquencer la vidéo requête à différents niveaux de précision. Le système modélise les relations de cause à effet entre les niveaux de descriptions ainsi que le déroulement temporel de la chirurgie à ces différents niveaux. Il a l'avantage d'être modulable : l'ajout d'un nouveau niveau de granularité est aisé, différents types d'observations peuvent être utilisés en entrée du modèle et il est possible de combiner différentes sources d'observations. Différentes configurations du modèle ont été évaluées avec deux niveaux de granularité (étape et phases) et des résultats encourageants ont été obtenus. Le modèle a notamment été validé avec une source d'observations fiable, fournie manuellement par les médecins : la présence des instruments chirurgicaux dans le champ de vu de la caméra. De très bonnes performances de reconnaissance ont été obtenues. Nous sommes donc capables de labelliser automatiquement la chirurgie à deux niveaux de description, dont un niveau précis (20 étapes chirurgicales) qui comprend des chevauchements et de nombreux ordonnancements possibles. La méthode est rapide et compatible avec une utilisation en temps réel. Des résultats encourageants ont également été obtenus avec l'analyse du mouvement dans la vidéo comme source d'observations. Dans le cas de l'utilisation de vidéos normalisées et d'un modèle un peu plus sophistiqué (retour des résultats du HMM vers le réseau bayésien), des performances satisfaisantes ont été obtenues. Ce système, bien que légèrement plus coûteux en temps de calcul, est également rapide et compatible avec une utilisation en temps réel. De nombreuses perspectives s'offrent avec ce système d'analyse multi-échelles de la chirurgie. L'une d'entre elle est le développement d'une méthode de détection et de reconnaissance automatique des instruments chirurgicaux. Nous pouvons également imaginer utiliser un modèle qui permette plus d'échanges entre le déroulement temporel et les relations entre les

niveaux, en s'inspirant des HMM hiérarchiques (hierarchical HMM) par exemple. Il serait également intéressant d'utiliser les données connues avant le démarrage de la chirurgie, tel que le type de cataracte (avancée ou non), le chirurgien qui opère, le type d'implant utilisé, etc… Afin d'adapter le modèle à la chirurgie qui va avoir lieu.

Pour conclure, la modélisation sous forme d'arbres n'a pas été concluante et nous a permis d'écarter cette piste. L'utilisation d'une modélisation multi-échelle est, quant à elle, concluante et pourra être adaptée par la suite pour la détection de déroulements anormaux et la génération d'alertes et de recommandations.

Chapitre V. Discussion générale

L'objectif de cette thèse était de mettre en place une méthode d'analyse automatique, en temps réel, de vidéos de chirurgie de la cataracte. Nous avons cherché à reconnaître à chaque instant de la chirurgie le geste chirurgical effectué. Plusieurs réponses à ce problème ont été apportées dans ce travail de thèse. Tour d'abord, par la conceptualisation du processus chirurgical à un niveau de précision élevé, grâce à une description multi-échelles de la cataracte. Puis par une méthode de recherche des cas les plus proches dans la base de données afin de catégoriser une séquence requête. Et enfin, par la création d'une modélisation statistique du processus chirurgical permettant d'apporter une information sur le déroulement probable de la chirurgie. Ces différents points ont permis d'aboutir à un séquençage automatique de la chirurgie de la cataracte, à un niveau de description fin, avec des bons taux de reconnaissance.

V.1 Description du processus chirurgical

Il existe plusieurs manières de décrire une même chirurgie, avec plus ou moins de précision. Il est nécessaire de trouver une description qui permette de décrire la chirurgie avec une précision adéquate et qui permette d'obtenir de bons taux de reconnaissance tout en ayant du sens d'un point de vue chirurgical. Dans des travaux précédents, une première description en tâches chirurgicales a montré quelques limites, notamment en termes de précision. C'est pourquoi nous avons mis au point, en collaboration avec un interne en chirurgie du service d'ophtalmologie du CHRU de Brest, une nouvelle description.

Cette nouvelle description est une description multi-échelle de la chirurgie. Pour cela, trois nouveaux niveaux de description ont été créés : les activités, les étapes et les phases. Cette nouvelle description est complète et pertinente d'un point de vue chirurgical. Elle répond tout d'abord à la nécessité d'analyser la chirurgie à un niveau de détail plus élevé. L'aspect multi-échelle permet d'entraîner des algorithmes qui fonctionnent mieux pour une analyse au niveau le plus fin. Cela permet également de choisir le niveau d'analyse souhaité et de générer des alertes et des recommandations propres à chaque niveau. Par exemple, l'analyse au niveau des activités pourra permettre d'apporter des recommandations sur le choix des outils, et l'analyse en étapes des recommandations sur le choix de la stratégie à adopter. Les alertes et les recommandations pourront ainsi être bien ciblées. Les diagrammes de transition construits pour chacun des niveaux de description permettent de se rendre compte de la grande variabilité de réalisation à chaque niveau, particulièrement pour les étapes et les phases. Cela montre la difficulté de travailler à ces niveaux de description fins, en termes de reconnaissance, mais également pour l'annotation manuelle de la base de données.

Seules 30 vidéos ont été annotées selon cette nouvelle description. Pour couvrir le plus grand nombre de cas opératoires, il sera nécessaire de continuer à enrichir cette base de données, de manière très conséquente. Cependant, l'annotation des vidéos est complexe et coûteuse en temps pour les chirurgiens qui doivent la réaliser : il existe un grand nombre d'activités différentes qui ne durent en général que quelques secondes. L'annotation des vidéos est donc un aspect clé

de ce travail. Ce problème a été partiellement résolu par la mise en place d'un système d'annotation qui permet d'annoter la vidéo seulement au niveau des activités. Les activités identiques, mais appartenant à des phases ou des étapes différentes, ont été différenciées au moment de l'annotation. Ainsi, les étapes et les phases sont déduites automatiquement lors de l'annotation. Cependant, l'annotation reste complexe et des erreurs d'annotations sont possibles. Pour limiter l'impact de ces erreurs, il serait envisageable d'identifier avec les chirurgiens un certain nombre de relations d'indépendances (de cooccurrences impossibles). Cela permettrait de construire un modèle statistique plus fiable. Il serait également envisageable d'utiliser le modèle statistique multi-échelles pour déduire les annotations des étapes et des activités à partir de la connaissance des labels « phases » et des observations. Cela permettrait de simplifier considérablement l'annotation des vidéos en n'annotant les vidéos qu'au niveau des phases.

V.2 Recherche des cas les plus proches

La recherche des cas les plus proches dans une base de données permet l'aide au diagnostic à partir de ces cas, ou dans notre contexte, une aide à la reconnaissance du geste chirurgical. Pour cela, il est nécessaire de représenter les vidéos par un vecteur de caractéristiques pour ensuite les comparer par une mesure de similitude adéquate.

Deux méthodes d'extraction de caractéristiques ont été évaluées : la construction d'histogramme de mouvements à partir du flux optique extrait entre deux images consécutives [49] et la construction d'histogrammes de mots visuels à partir de descripteurs STIP [68]. Afin d'affiner l'extraction de ces caractéristiques de mouvement, nous avons évalué une normalisation des vidéos via trois sortes de prétraitements des images : le recalage de l'iris au centre de l'image, la mise à l'échelle des vidéos (tailles d'iris identiques) et la sélection d'une région d'intérêt (ROI). Une mesure de similitude proposée par Piciarelli et al. [44], alternative à la distance DTW a été utilisée. L'avantage de cette mesure de similitude est tout d'abord sa simplicité, mais également son mode de calcul « à la volée », c'est-à-dire que la mesure est mise à jour à chaque acquisition d'une nouvelle image. Ainsi, contrairement à la mesure DTW, elle ne nécessite pas d'attendre la fin de l'enregistrement de la séquence.

Des résultats satisfaisants ont été obtenus compte tenu de la simplicité de l'algorithme mis en place dans ce travail. La méthode répond aux contraintes de temps réel. Les meilleurs résultats en termes de performances de reconnaissance ont été obtenus avec les histogrammes de mots visuels comme signatures visuelles des séquences avec une aire moyenne sous la courbe ROC de $A_z = 0{,}826$ (et $A_z = 0{,}728$ pour les histogrammes de mouvement). Cependant, la construction de ces signatures n'est pas compatible avec une utilisation en temps réel, avec moins d'une image traitée par seconde. La normalisation des vidéos, compatible avec une utilisation temps réel, a permis d'améliorer les performances de reconnaissances ($A_z = 0{,}837$ pour les histogrammes de mots visuels et $A_z = 0{,}794$ pour les histogrammes de mouvements). Néanmoins, certains aspects de cette normalisation peuvent être améliorés. C'est le cas du choix du rayon de la région d'intérêt sélectionnée. Il serait intéressant d'étudier l'impact du choix de ce rayon sur les performances de reconnaissance. L'utilisation de la mesure de similitude de Piciarelli [44] s'est révélée intéressante. Nous avons constaté que les performances de reconnaissance sont directement liées au choix du paramètre δ fixant la vitesse d'accroissement de la fenêtre glissante. Ce paramètre a

été fixé à $\delta = 0{,}1$, mais pour certaines tâches chirurgicales telles que l'hydrodissection, les performances de reconnaissance sont supérieures avec une valeur de δ plus élevée. Les travaux futurs pourraient chercher des paramètres spécifiques à la tâche recherchée.

Enfin, cette méthode n'a pas été évaluée avec le modèle multi-échelle du processus chirurgical. Il serait néanmoins intéressant d'évaluer la combinaison des deux méthodes. Cela nécessiterait néanmoins d'utiliser une méthode de structuration de l'espace de recherche pour gérer le grand nombre de sous-séquences de la base d'apprentissage.

V.3 Modélisation statistique du processus chirurgical

La reconnaissance des différentes étapes de la chirurgie, à chaque instant, et donc l'annotation automatique des vidéos est la condition sine qua non du suivi en temps réel des chirurgies pour apporter une aide per-opératoire aux chirurgiens. La modélisation statistique du processus chirurgical va permettre d'utiliser la connaissance des déroulements probables de la chirurgie, appris à partir de la base de données, pour apporter une information contextuelle lors de l'analyse d'une vidéo requête.

On l'a constaté, bien que la chirurgie de la cataracte soit une chirurgie très reproductible, il existe néanmoins de nombreux déroulements possibles. Ces déroulements sont d'autant plus complexes qu'on utilise un niveau de description fin. Le nombre de vidéos annotées disponibles dans la base de données est donc une des principales limites à la mise en œuvre de cette méthode. Nous l'avons néanmoins évaluée à partir de deux modèles statistiques. Nous avons montré que le premier modèle, un arbre des déroulements possibles de la chirurgie construit via la méthode de Piciarelli et al. [44], ne fonctionne pas : nous l'avons donc écarté. Le second modèle évalué s'appuie sur l'aspect multi-échelle de la chirurgie de la cataracte. Il a montré des résultats satisfaisants avec l'utilisation de deux niveaux de description : étapes et phases. Le modèle a été validé selon différentes configurations. De meilleurs résultats ont été obtenus pour la description en phases ($A_z = 0{,}691$ avec les CRF et $A_z = 0{,}674$ avec les HMM) que pour la description en étapes ($A_z = 0{,}828$ avec les CRF et $A_z = 0{,}812$ avec les HMM). Les performances de reconnaissances peuvent être améliorées avec la normalisation des vidéos ($A_z = 0{,}721$ pour les étapes et $A_z = 0{,}832$ pour les phases, avec les HMM). Des très bons résultats ont été obtenus avec la présence des instruments comme source d'observation, particulièrement avec la modélisation du déroulement temporel par des CRF ($A_z = 0{,}980$ pour les étapes et $A_z = 0{,}986$ pour les phases).

Notons que le second modèle n'a pas été évalué avec le niveau de granularité le plus fin : les activités. Or, c'est à ce niveau de description que nous souhaitons être capable d'analyser la chirurgie. Cependant, plus la description est fine, plus la reconnaissance est complexe. En particulier, dans le cas des activités, nous n'avons pas réussi à estimer de manière satisfaisante les probabilités conditionnelles au sein du réseau bayésien et des HMM, et donc la structure de ces graphes qui en découle. Le problème vient du déséquilibre entre le faible nombre d'exemples d'apprentissage et le nombre élevé de degrés de liberté. Pour résoudre ce problème, il faudrait soit intégrer de la connaissance expert, soit trouver un processus d'apprentissage plus robuste.

Il ressort des évaluations qu'une limite du modèle est le manque de communication entre le réseau bayésien et les modèles de Markov. En ne regardant que les phases et les étapes, les meilleurs résultats ont en effet été obtenus pour les phases suite à l'inférence des HMM, ce qui montre

qu'il serait intéressant de pouvoir retransmettre aisément cette information au réseau bayésien. Il serait intéressant d'évaluer un autre type de modèle tel que les « hierarchical HMM » qui permettent d'introduire plusieurs niveaux de granularité dans un modèle de Markov caché. Chaque état du modèle est alors lui-même un HMM. Il apparaît également que les CRF donnent de meilleurs résultats que les HMM, il serait alors intéressant de poursuivre les expérimentations avec ce type de modélisation statistique du processus chirurgical. Une autre limite du modèle vient de la mesure de distance utilisée pour comparer des sous-séquences vidéo. Celle-ci est utilisée pour la recherche des plus proches voisins qui fournit des preuves au modèle à partir du mouvement extrait dans la vidéo. Or, cette recherche a été réalisée via la méthode ANN[10]. Cette méthode suppose que les distances sont mesurées avec une distance de Minkowski (distance de Manhattan, distance euclidienne, etc…). Nous avons utilisé une distance euclidienne. Or nous avons vu dans le Chapitre III que cette mesure de similitude entre deux signatures visuelles n'était pas idéale dans le cas de nos signatures. De plus la méthode requiert un vecteur de caractéristiques par séquence pour la recherche de plus proche voisin. C'est pourquoi nous avons utilisé un histogramme du cumul du mouvement au sein de la séquence. Cependant, l'utilisation d'un vecteur de caractéristiques prenant mieux en compte les variations de mouvement au sein de la séquence pourrait s'avérer plus judicieux. Il serait intéressant d'évaluer la méthode avec le vecteur de caractéristiques utilisé par Quellec et al. [49], par exemple. Enfin, de très bons résultats ont été obtenus avec la présence des instruments chirurgicaux comme source d'observations. Cette information est actuellement fournie via les annotations manuelles des chirurgiens. L'automatisation de la reconnaissance des instruments s'avère donc être une tâche très importante à réaliser pour le suivi automatique temps réel des chirurgies.

10 http://www.cs.umd.edu/~mount/ANN/

Conclusion

De plus en plus, les données de chirurgies et notamment les vidéos de contrôle sont stockées dans des archives médicales numériques. Ce travail de thèse avait pour objectif la réutilisation de ces archives pour l'aide à la chirurgie en temps réel. Les vidéos de chirurgies précédemment enregistrées, contenues dans les archives médicales numériques, sont peu exploitées car difficilement consultables par les médecins. Or ces vidéos constituent une source importante de connaissances, à travers des exemples de chirurgiens plus expérimentés, des cas particuliers, des situations délicates, etc. En particulier, ces archives sont difficilement accessibles lorsque le chirurgien en a le plus besoin : lorsqu'il opère. Dans cette thèse, nous nous sommes intéressés à la reconnaissance automatique du geste chirurgical : si nous sommes capables de reconnaître à chaque instant ce que fait le chirurgien, alors nous pouvons déterminer automatiquement l'information dont il peut avoir besoin : que font ses collègues lorsqu'ils sont dans une situation similaire ? Est-ce que cette situation a déjà mené à des complications ? Nous avons donc cherché dans cette thèse à analyser automatiquement une nouvelle vidéo de chirurgie pour générer par la suite des alertes et des recommandations adaptées. Nous nous sommes particulièrement intéressés à la reconnaissance automatique du geste chirurgical, en temps réel. Cela revient à segmenter automatiquement une nouvelle vidéo de chirurgie en gestes chirurgicaux.

A ce jour, il n'existe pas de bases de données communes de vidéos chirurgicales permettant d'entraîner et de tester de tels outils. Le LaTIM, avec le service d'ophtalmologie du CHRU de Brest, a alors construit sa propre base de vidéos chirurgicales. Nous nous sommes concentrés sur la chirurgie de la cataracte, l'une des chirurgies les plus pratiquées. Grâce à un travail réalisé en collaboration avec un interne en chirurgie du service d'ophtalmologie du CHRU de Brest, une description multi-échelle de la chirurgie a alors été mise en place. Pour cela, trois nouveaux niveaux de description ont été créés : activités (le plus précis), étapes et phases (le plus grossier). Cette description permet de décrire de façon complète la chirurgie de la cataracte. Les alertes et les recommandations pourront ainsi être bien ciblées.

Nous avons ensuite cherché à reconnaître automatiquement le geste chirurgical effectué au sein d'une portion de vidéo. Nous nous sommes appuyés pour cela sur le principe de la recherche par le contenu : en cherchant les portions de vidéos les plus proches dans la base de données, nous pouvons déterminer par analogie le geste le plus probable. Pour cela, nous nous sommes appuyés sur une base de séquences vidéo ou chaque séquence représente une tâche chirurgicale réalisée lors d'une chirurgie de la cataracte. La recherche des cas similaires comporte deux points clés : la caractérisation des vidéos par des signatures visuelles et la mesure de similitude entre le cas requête et les cas de la base de données. Nous avons choisi de caractériser nos vidéos en analysant le mouvement qu'elles contiennent. Différentes signatures numériques ont été évaluées pour caractériser le mouvement, après avoir normalisé la taille et le centre de l'iris dans les vidéos, afin de ne garder que les mouvements du chirurgien. La mesure de similitude évaluée est issue des méthodes d'analyse de vidéosurveillance. Cette mesure de similitude a été proposée par Piciarelli et al. [44] et est une alternative à la distance DTW. Elle a été adaptée à la comparaison de séquences vidéo médicales. Une aire moyenne sous la courbe ROC $A_z = 0{,}728$ a été obtenue avec une signature simple calculable en temps réel et ne aire moyenne $A_z = 0{,}826$ a été obtenue

avec une signature plus évoluée mais calculable uniquement hors ligne. Les résultats obtenus sont satisfaisants compte tenu de la simplicité et de la rapidité du système.

Enfin, nous avons cherché à segmenter une vidéo de chirurgie complète, en temps réel. Pour cela, des observations sont générées à pas de temps fixes, puis interprétées par un modèle statistique du déroulement de la chirurgie. Plusieurs modèles probabilistes, à base d'arbres et de graphes en général, ont ainsi été proposés et évalués. Le modèle que je préconise utilise l'aspect multi-échelles de la description et modélise, par des graphes probabilistes, le processus chirurgical en utilisant à la fois les relations qui existent entre les différents niveaux de description et les relations temporelles qui existent entre les labels de chacun des niveaux. Ce modèle a été évalué avec deux niveaux de granularité : les étapes et les phases. Pour la modélisation des relations entre les niveaux de description, un réseau bayésien a été utilisé. Pour la modélisation du déroulement temporel de la chirurgie, un modèle de Markov caché (HMM) ou un champ markovien conditionnel (CRF) a été utilisé pour chacun des niveaux de description. Deux types d'observations ont été évalués : l'information de présence des instruments dans le champ de vue de la caméra (obtenue de manière non automatique) et les résultats d'une recherche des cas les plus proches via l'analyse du mouvement dans la vidéo (obtenus de manière automatique). Des résultats satisfaisants ont été obtenus en nous appuyant sur l'analyse du mouvement et la recherche des cas les plus proches en entrée du système: une aire moyenne $A_z = 0{,}721$ a été obtenue pour les étapes, une aire moyenne $A_z = 0{,}832$ a été obtenue pour les phases. Nous avons ainsi été capables d'analyser la chirurgie de la cataracte à un niveau de description relativement fin, avec des taux de reconnaissance satisfaisants, tout en respectant la contrainte de temps réel. Des très bons résultats ont été obtenus en nous appuyant sur la présence des instruments chirurgicaux, issue d'une segmentation manuelle, en entrée du système. C'est particulièrement vrai avec la modélisation du déroulement temporel par des CRF ($A_z = 0{,}980$ pour les étapes et $A_z = 0{,}986$ pour les phases). Ces résultats motivent la reconnaissance automatique des instruments chirurgicaux.

La reconnaissance automatique des instruments chirurgicaux en cours d'utilisation peut se faire de deux manières : via l'utilisation de méthodes de détection et de reconnaissance automatique ou via l'utilisation de méthodes non basées sur l'analyse temps réel des flux vidéos disponibles, par exemple des radio-étiquettes (RFID). Pour des raisons d'asepsie, il est difficile d'intégrer des éléments externes, passifs ou actifs, aux instruments chirurgicaux. En termes d'asepsie, la détection et reconnaissance automatique des instruments chirurgicaux, à partir du flux vidéo, semblent donc la solution la plus facile à réaliser. Mais de fait, il s'agit d'une tâche complexe sur le plan méthodologique, et la méthode doit respecter les contraintes de temps réel. Au niveau images, notons que seule la tête des instruments est visible dans le champ de vue de la caméra, et celle-ci est souvent déformée par son entrée dans l'œil. De plus certains instruments se ressemblent et sont difficilement différentiables. Un sujet de thèse a débuté sur ce sujet au sein de l'équipe. L'idée développée dans cette thèse est de s'appuyer sur un second flux vidéo qui filme la table des instruments chirurgicaux pour obtenir une information supplémentaire. L'avantage de cette approche est que les instruments chirurgicaux sont entièrement visibles sur la table, et que la sortie ou l'entrée d'un instrument de la table implique un changement probable des instruments présents dans la scène chirurgicale.

Le travail développé dans cette thèse est une première base pour la mise en place future de la détection d'événements anormaux ou pouvant donner lieu à des complications. Il est alors nécessaire d'être capable de détecter des événements (ou des séquences d'événements) annonciateurs d'une complication. Les méthodes de fouille de données utilisées pour la modélisation statistique du processus chirurgical peuvent être adaptées à la reconnaissance automatique

d'événements anormaux ou d'événements pouvant conduire à des situations anormales. Il sera ensuite nécessaire de définir des alertes et des recommandations adaptées. A partir des actions effectuées par des médecins expérimentés dans des situations similaires, on cherchera à définir des préconisations, voire déclencher des alertes et proposer des actions. Ces alertes et recommandations devront être amenées de façon judicieuse afin de ne pas perturber le travail du chirurgien. Ce point devra donc être l'objet d'une profonde réflexion avec les praticiens.

Les applications de ce travail de thèse sont nombreuses. La première application est, bien sûr, dans le sujet de ma thèse : l'aide à la prise de décision lors de la pratique chirurgicale. Cette aide peut être un véritable atout pour les jeunes chirurgiens en apprentissage. Cet outil peut par exemple être intégré aux simulateurs de gestes chirurgicaux, pour apporter des indications aux chirurgiens qui s'entraînent sur des gestes ciblés. L'outil peut également, dans cette situation, être utilisé pour évaluer la qualité du geste chirurgical réalisé ou évaluer les compétences du jeune chirurgien et quantifier sa progression. Des conseils adaptés peuvent alors lui être apportés. Cette aide peut ensuite se poursuivre lors de la pratique de ces jeunes chirurgiens en situation réelle. Il sera alors nécessaire de réfléchir à la manière optimale d'apporter les alertes et les recommandations nécessaires. Ce travail pourrait aussi être une piste intéressante pour le contrôle des chirurgies robotisées. L'outil développé peut être adapté pour générer des commandes adaptées ou des conditions d'arrêts. Dans le cas des robots manipulés, tels que le robot Da Vinci, il pourrait être également utilisé pour générer des indications au chirurgien qui manipule. Nous pouvons également imaginer utiliser le modèle multi-échelles mis en place pour l'archivage et l'annotation automatique ou semi-automatique de vidéos. La méthode a été développée avec des contraintes de temps réel. Cependant, pour l'indexation d'archives, il est possible d'utiliser des sources d'observations plus complexes, même si elles sont plus coûteuses en temps de calcul. Nous pouvons ainsi espérer améliorer les performances de reconnaissance.

En conclusion, ce travail est un premier pas vers l'interrogation de bases de vidéos chirurgicales en cours de chirurgie. Les applications potentielles d'un tel outil sont nombreuses et vont permettre d'améliorer la formation des médecins et la sécurité des patients.

Bibliographie

[1] Shu-Hsien Liao, Pei-Hui Chu, and Pei-Yuan Hsiao, "Data mining techniques and applications--A decade review from 2000 to 2011," *Expert Systems with Applications*, vol. 39, no. 12, pp. 11303-11311, 2012.

[2] Agnar Aamodt and Enric Plaza, "Case-based reasoning: Foundational issues, methodological variations, and system approaches," *AI communications*, vol. 7, no. 1, pp. 39-59, 1994.

[3] Mehwish Rehman, Muhammad Iqbal, Muhammad Sharif, and Mudassar Raza, "Content based image retrieval: survey," *World Applied Sciences Journal*, vol. 19, no. 3, pp. 404-412, 2012.

[4] Etienne Decencière et al., "TeleOphta: Machine learning and image processing methods for teleophthalmology," *IRBM*, vol. 34, no. 2, pp. 196-203, 2013.

[5] Henning Müller, Nicolas Michoux, David Bandon, and Antoine Geissbuhler, "A review of content-based image retrieval systems in medical applications—clinical benefits and future directions," *International journal of medical informatics*, vol. 73, no. 1, pp. 1-23, 2004.

[6] Nianhua Xie, Li Li Xianglin Zeng Weiming Hu and Stephen Maybank, "A Survey on Visual Content-Based Video Indexing and Retrieval," *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS*, vol. 41, no. 6, pp. 797-819, November 2011.

[7] Yaqin Li et al., "A health insurance portability and accountability act--compliant ocular telehealth network for the remote diagnosis and management of diabetic retinopathy," *Telemedicine and e-Health*, vol. 17, no. 8, pp. 627-634, 2011.

[8] Carlos M Oliveira, Luis M Cristóvão, Maria Luisa Ribeiro, and JR Abreu, "Improved automated screening of diabetic retinopathy," *Ophthalmologica*, vol. 226, no. 4, pp. 191-197, 2011.

[9] JR Ordóñez, G Cazuguel, J Puentes, B Solaiman, and C Roux, "Indexation d'images médicales basée sur les informations spectrale et spatiale extraites de JPEG-2000".

[10] Gwénolé Quellec et al., "Optimal wavelet transform for the detection of microaneurysms in retina photographs," *Medical Imaging, IEEE Transactions on*, vol. 27, no. 9, pp. 1230-1241, 2008.

[11] Gwénolé Quellec, Mathieu Lamard, Guy Cazuguel, Béatrice Cochener, and Christian Roux, "Adaptive nonseparable wavelet transform via lifting and its application to content-

based image retrieval," *Image Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 25-35, 2010.

[12] Gwénolé Quellec, Mathieu Lamard, Guy Cazuguel, Béatrice Cochener, and Christian Roux, "Fast wavelet-based image characterization for highly adaptive image retrieval," *Image Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 1613-1623, 2012.

[13] Gwénolé Quellec, Mathieu Lamard, Guy Cazuguel, Béatrice Cochener, and Christian Roux, "Wavelet optimization for content-based image retrieval in medical databases," *Medical image analysis*, vol. 14, no. 2, pp. 227-241, 2010.

[14] Said Jai-Andaloussi et al., "content based medical image retrieval: use of Generalized Gaussian Density to model BEMD's IMF," in *World Congress on Medical Physics and Biomedical Engineering, September 7-12, 2009, Munich, Germany*, 2010, pp. 1249-1252.

[15] Said Jai Andaloussi, "Indexation de l'information médicale. Application à la recherche d'images et de vidéos par le contenu," Ph.D. dissertation 2010.

[16] Gwénolé Quellec, "Indexation et fusion multimodale pour la recherche d'information par le contenu. Application aux bases de données d'images médicales.," Université de Rennes I, École doctorale Matisse, Ph.D. dissertation 2008.

[17] Gwénolé Quellec, Mathieu Lamard, Guy Cazuguel, Christian Roux, and Béatrice Cochener, "Case retrieval in medical databases by fusing heterogeneous information," *Medical Imaging, IEEE Transactions on*, vol. 30, no. 1, pp. 108-118, 2011.

[18] Sean R Stanek, Wallapak Tavanapong, Johnny Wong, Jung Hwan Oh, and Piet C De Groen, "Automatic real-time detection of endoscopic procedures using temporal features," *Computer Methods and Programs in Biomedicine*, vol. Volume 108, Issue 2, pp. 524-535, 2012.

[19] Danyu Liu, Wallapak Tavanapong Johnny Wong JungHwan Oh Yu Cao and Piet C. de Groen, "Computer-Aided Detection of Diagnostic and Therapeutic Operations in Colonoscopy Videos," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 7, pp. 1268-1279, July 2007.

[20] Andru Putra Twinanda, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy, "Classification approach for automatic laparoscopic video database organization," *International journal of computer assisted radiology and surgery*, pp. 1-12, 2015.

[21] Florent Lalys, David Bouget, Laurent Riffaud, and Pierre Jannin, "Automatic knowledge-based recognition of low-level tasks in ophthalmological procedures," *International journal of computer assisted radiology and surgery*, vol. 8, no. 1, pp. 39-49, 2013.

[22] Florent Lalys, Laurent Riffaud, David Bouget, and Pierre Jannin, "A framework for the recognition of high-level surgical tasks from video images for cataract surgeries," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 966-976, 2012.

[23] Florent LALYS, "Automatic recognition of low-level and high-level surgical tasks in the Operating Room from video images," Vie-Agro-Santé, Ph.D. dissertation 2012.

[24] Tobias Blum, Hubertus Feußner, and Nassir Navab, "Modeling and segmentation of surgical workflow from laparoscopic video," in *Medical Image Computing and Computer-Assisted Intervention--MICCAI 2010*, 2010, pp. 400-407.

[25] Nicolas Padoy et al., "Statistical modeling and recognition of surgical workflow," *Medical Image Analysis*, vol. 16, no. 3, pp. 632-641, April 2012.

[26] Germain Forestier, Laurent Riffaud, and Pierre Jannin, "Automatic phase prediction from low-level surgical activities," *International journal of computer assisted radiology and surgery*, pp. 1-9, 2015.

[27] Michal Mackiewicz, Jeff Berens, and Mark Fisher, "Wireless Capsule Endoscopy Color Video Segmentation," *IEEE Transaction on Medical Imaging*, vol. 27, no. 12, pp. 1769-1781, December 2008.

[28] Mark Fisher and Michal Mackiewicz, "Colour Image Analysis of Wireless Capsule Endoscopy Video: A Review," in *Color Medical Image Analysis*.: Springer, 2013, pp. 129-144.

[29] Benjamin Béjar Haro, Luca Zappella, and René Vidal, "Surgical gesture classification from video data," in *Medical Image Computing and Computer-Assisted Intervention--MICCAI 2012*.: Springer, 2012, pp. 34-41.

[30] Lingling Tao, Luca Zappella, Gregory D Hager, and René Vidal, "Surgical gesture segmentation and recognition," in *Medical Image Computing and Computer-Assisted Intervention--MICCAI 2013*.: Springer, 2013, pp. 339-346.

[31] Luca Zappella, Benjamin Béjar, Gregory Hager, and René Vidal, "Surgical gesture classification from video and kinematic data," *Medical image analysis*, vol. 17, no. 7, pp. 732-745, 2013.

[32] Barbara André, Tom Vercauteren, Anna M Buchner, Michael B Wallace, and Nicholas Ayache, "Learning semantic and visual similarity for endomicroscopy video retrieval," *IEEE Transactions on Medical Imaging*, vol. 31, no. 6, pp. 1276-1288, 2012.

[33] Ignacio Oropesa et al., "EVA: Laparoscopic Instrument Tracking Based on Endoscopic Video Analysis for Psychomotor Skills Assessment," *Surgical endoscopy*, vol. 27, no. 3, pp. 1029-1039, 2013.

[34] Julian Leong et al., "HMM assessment of quality of movement trajectory in laparoscopic surgery," , 2006, pp. 752-759.

[35] Takahisa Suzuki et al., "An evaluation of the endoscopic surgical skills assessment using a video analysis software program," *Surgical endoscopy*, pp. 1-5, 2014.

[36] Carol E Reiley and Gregory D Hager, "Task versus subtask surgical skill evaluation of robotic minimally invasive surgery," in *Medical Image Computing and Computer-Assisted Intervention--MICCAI 2009*.: Springer, 2009, pp. 435-442.

[37] Carol E Reiley et al., "Automatic recognition of surgical motions using statistical modeling for capturing variability," *Studies in health technology and informatics*, vol. 132, p. 396, 2008.

[38] Carol E Reiley, Henry C Lin, David D Yuh, and Gregory D Hager, "Review of methods for objective surgical skill evaluation," *Surgical endoscopy*, vol. 25, no. 2, pp. 356-366, 2011.

[39] Lingling Tao, Ehsan Elhamifar, Sanjeev Khudanpur, Gregory D Hager, and René Vidal, "Sparse hidden markov models for surgical gesture classification and skill evaluation," in *Information Processing in Computer-Assisted Interventions*.: Springer, 2012, pp. 167-177.

[40] Charles Garraud et al., "An Ontology-based Software Suite for the Analysis of Surgical Process Model," 2014.

[41] Beenish Bhatia, Tim Oates, Yan Xiao, and Peter Hu, "Real-Time Identification of Operating Room State from Video," in *Proc. of Innovative Applications of Artificial Intelligence (IAAI)*, 2007, pp. 1761–1766.

[42] Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi, "Trajectory pattern mining," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2007, pp. 330-339.

[43] Gerardo Castanon, Greg Castañón Venkatesh Saligrama, and Andre Louis and Jodoin, Pierre-Marc Caron, "Real-Time Activity Search of Surveillance Video," in *IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, 2012, pp. 246-251.

[44] Claudio Piciarelli and Gian Luca Foresti, "On-line trajectory clustering for anomalous events detection," *Pattern Recognition Letters*, vol. 27, no. 15, pp. 1835-1842, 2006.

[45] Javier Acevedo-Rodríguez, Saturnino Maldonado-Bascón, Roberto López-Sastre, Pedro Gil-Jiménez, and Antonio Fernández-Caballero, "Clustering of Trajectories in Video Surveillance Using Growing Neural Gas," in *Foundations on Natural and Artificial Computation*, 2011, pp. 461-470.

[46] Timothy Hospedales, Shaogang Gong, and Tao Xiang, "Video behaviour mining using a dynamic topic model," *International journal of computer vision*, vol. 98, no. 3, pp. 303-323, 2012.

[47] Gwénolé Quellec, Mathieu Lamard Guy Cazuguel Béatrice Cochener Christian Roux Zakarya Droueche, Computer-Aided Retinal Surgery using Data from the Video Compressed Stream, 2012.

[48] Zakarya DROUECHE, "Fouille de séquence d'images médicales. Application en chirurgie mini-invasive augmentée," Télécom Bretagne, Ph.D. dissertation 2012.

[49] Gwénolé Quellec et al., "Real-time recognition of surgical tasks in eye surgery videos," *Medical image analysis*, vol. 18, no. 3, pp. 579-590, 2014.

[50] Gwénolé Quellec, Mathieu Lamard, Béatrice Cochener, and Guy Cazuguel, "Real-Time Task Recognition in Cataract Surgery Videos using Adaptive Spatiotemporal Polynomials," *IEEE Transactions on Medical Imaging*, vol. 34, no. 4, 2015.

[51] Gwénolé Quellec, Mathieu Lamard, Béatrice Cochener, and Guy Cazuguel, "Real-time segmentation and recognition of surgical tasks in cataract surgery videos," *IEEE Transactions on Medical Imaging*, vol. 33, no. 12, 2014.

[52] P. Indyk A. Gionis and R. Motwani, "Similarity Search in High Dimensions via Hashing," in *Proceedings of the 25th Very Large Data Bases Conference*, 1999, pp. 518-529.

[53] Loïc Paulevé, Hervé Jégou, and Laurent Amsaleg, "Locality sensitive hashing: A comparison of hash function types and querying mechanisms," *Pattern Recognition Letters*, vol. 31, no. 11, pp. 1348-1358, 2010.

[54] Mani Malek Esmaeili, Rabab Kreidieh Ward, and Mehrdad Fatourechi, "A Fast Approximate Nearest Neighbor Search Algorithm in the Hamming Space," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 12, pp. 2481-2488, December 2012.

[55] Florent Lalys and Pierre Jannin, "Surgical process modelling: a review," *International journal of computer assisted radiology and surgery*, pp. 1-17, 2013.

[56] Barbara André, Tom Vercauteren, Anna M Buchner, Michael B Wallace, and Nicholas Ayache, "Endomicroscopic video retrieval using mosaicing and visualwords," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2010, pp. 1419-1422.

[57] DH Martiano, K Charrière, M Lamard, and B Cochener, "Indexing of cataract surgery video by content based video retrieval," *Acta Ophthalmologica*, vol. 92, no. s253, pp. 0-0, 2014.

[58] Giovanni Fusco, Nicoletta Noceti, and Francesca Odone, "Combining retrieval and classification for real-time face recognition," in *IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, September 2012, p. 276281.

[59] Ja-Hwung Su, Yu-Ting Huang, Hsin-Ho Yeh, and Vincent S Tseng, "Effective content-based video retrieval using pattern-indexing and matching techniques," *Expert Systems with Applications*, vol. 37, no. 7, pp. 5068-5085, July 2010.

[60] Gwénolé Quellec et al., "Real-Time Retrieval of Similar Videos with Application to Computer-Aided Retinal Surgery," in *Engineering in Medicine and Biology Society,EMBC, 2011 Annual International Conference of the IEEE*, 2011, pp. 4465-4468.

[61] B. Ramamurthy and K.R.Chandran, "Content Based Medical Image Retrieval with Texture Content Using Gray Level Co-occurrence Matrix and K-Means Clustering Algorithms," *Journal of Computer Science*, vol. 8, no. 7, pp. 1070-1076, 2012.

[62] Gwénolé Quellec et al., "A polynomial model of surgical gestures for real-time retrieval of surgery videos," in *Medical Content-Based Retrieval for Clinical Decision Support*.: Springer, 2013, pp. 10-20.

[63] Chuen-Horng Lin, Der-Chen Huang, Yung-Kuan Chan, Kai-Hung Chen, and Yen-Jen Chang, "Fast color-spatial feature based image retrieval methods," *Expert Systems with Applications*, vol. 38, no. 9, pp. 11412-11420, September 2011.

[64] Xinbo Gao, Xuelong Li, Jun Feng, and Dacheng Tao, "Shot-based video retrieval with optical flow tensor and HMMs," *Pattern regignition Letters*, vol. 30, no. 2, pp. 140-147, January 2009.

[65] Faisal and Khokhar, Ashfaq and Schonfeld, Dan and others Bashir, "Real-Time Motion Trajectory-Based Indexing and Retrieval of Video Sequences," *IEEE TRANSACTIONS ON MULTIMEDIA*, vol. 9, no. 1, pp. 58-65, January 2007.

[66] David M Blei, Andrew Y Ng, and Michael I Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[67] David G Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2, 1999, pp. 1150-1157.

[68] Ivan Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107-123, 2005.

[69] Chris Harris and Mike Stephens, "A combined corner and edge detector.," in *Alvey vision conference*, vol. 15, 1988, p. 50.

[70] Donald J Berndt and James Clifford, "Using Dynamic Time Warping to Find Patterns in Time Series.," in *KDD workshop*, vol. 10, 1994, pp. 359-370.

[71] Ghazi Al-Naymat, Sanjay Chawla, and Javid Taheri, "SparseDTW: a novel approach to speed up dynamic time warping," in *Proceedings of the Eighth Australasian Data Mining Conference-Volume 101*, 2009, pp. 117-127.

[72] Eamonn Keogh and Chotirat Ann Ratanamahatana, "Exact indexing of dynamic time warping," *Knowledge and information systems*, vol. 7, no. 3, pp. 358-386, 2005.

[73] Li Gao, Zhu Li, and Aggelos Katsaggelos, "An Efficient Video Indexing and Retrieval Algorithm Using the Luminance Field Trajectory Modeling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 10, pp. 1566-1570, October 2009.

[74] Michail Vlachos, Marios Hadjieleftheriou, Dimitrios Gunopulos, and Eamonn Keogh, "Indexing multi-dimensional time-series with support for multiple distance measures," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003, pp. 216-225.

[75] Bruce D Lucas and Takeo Kanade, "An iterative image registration technique with an application to stereo vision.," in *IJCAI*, vol. 81, 1981, pp. 674-679.

[76] Gwenole Quellec, Katia Charriere, Mathieu Lamard, Beatrice Cochener, and Guy Cazuguel, "Normalizing videos of anterior eye segment surgeries," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, 2014, pp. 122-125.

[77] Theo Gasser and Kongming Wang, "Alignment of curves by dynamic time warping," *The Annals of Statistics*, vol. 25, no. 3, pp. 1251-1276, 1997.

[78] Forney Jr and G David, "The viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268-278, 1973.

[79] John Lafferty, Andrew McCallum, and Fernando CN Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[80] Patrick Naïm, Pierre-Henri Wuillemin, Philippe Leray, Olivier Pourret, and Anna Becker, *Réseaux bayésiens*.: Editions Eyrolles, 2011.

[81] Luis M De Campos, "A scoring function for learning bayesian networks based on mutual information and conditional independence tests," *The Journal of Machine Learning Research*, vol. 7, pp. 2149-2187, 2006.

[82] Thomas Lavergne, Olivier Cappè, and François Yvon, "Practical Very Large Scale

[83] Judea Pearl and Stuart Russell, *Bayesian networks*.: Computer Science Department, University of California, 1998.

[84] Finn V Jensen, Steffen L Lauritzen, and Kristian G Olesen, "Bayesian updating in causal probabilistic networks by local computations," *Computational statistics quarterly*, vol. 4, pp. 269-282, 1990.

[85] Radford M Neal, "Probabilistic inference using Markov chain Monte Carlo methods," 1993.

[86] Fei Sha and Fernando Pereira, "Shallow parsing with conditional random fields," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, 2003, pp. 134-141.

[87] Yu Cao, Shih-Hsi Liu, Ming Li, Sung Baang, and Sanqing Hu, "Medical video event classification using shared features," in *Tenth IEEE International Symposium on Multimedia (ISM).*, 2008, pp. 266-273.

Publications

K. Charrière, G. Quellec, D. Martiano, M. Lamard, G. Cazuguel, G. Coatrieux, and B. Cochener, “A multilevel statistical model for the automatic analysis of cataract surgeries,” in MICCAI Workshop on Medical Content-based Retrieval for Clinical Decision Support, 2015.

K. Charrière, G. Quellec, M. Lamard, G. Coatrieux, B. Cochener, and G. Cazuguel, “Automated surgical step recognition in normalized cataract surgery videos,” in 36th Annual International Conference of the IEEE Ingineering in Medicine and Biology Society (EMBC). IEEE, 2014, pp. 4647–4650.

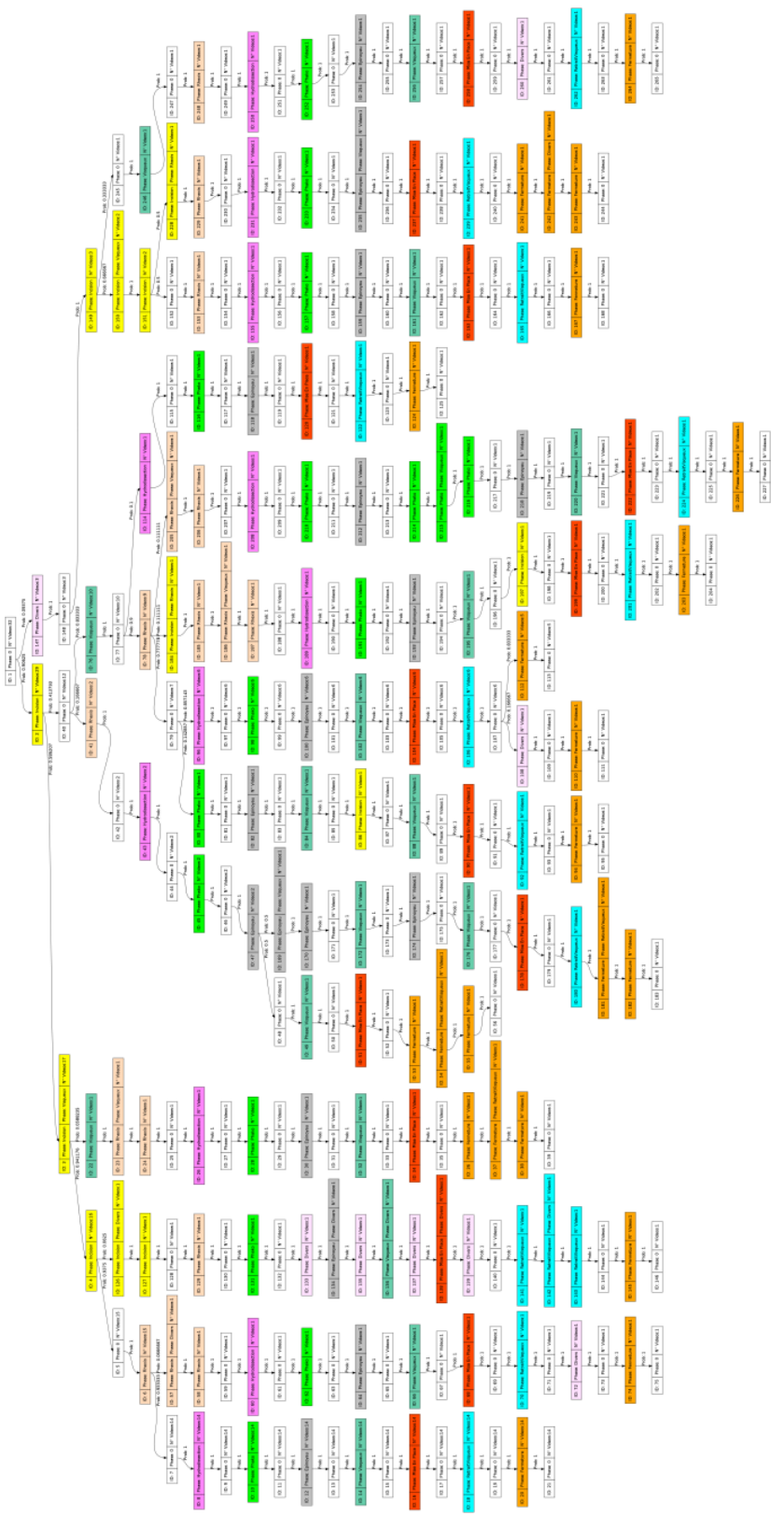
G. Quellec, K. Charrière, M. Lamard, Z. Droueche, C. Roux, B. Cochener, and G. Cazuguel, “Real-time recognition of surgical tasks in eye surgery videos,” Medical image analysis, vol. 18, no. 3, pp. 579–590, 2014.

G. Quellec, K. Charrière, M. Lamard, B. Cochener, and G. Cazuguel, “Normalizing videos of anterior eye segment surgeries,” in Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE. IEEE, 2014, pp. 122–125.

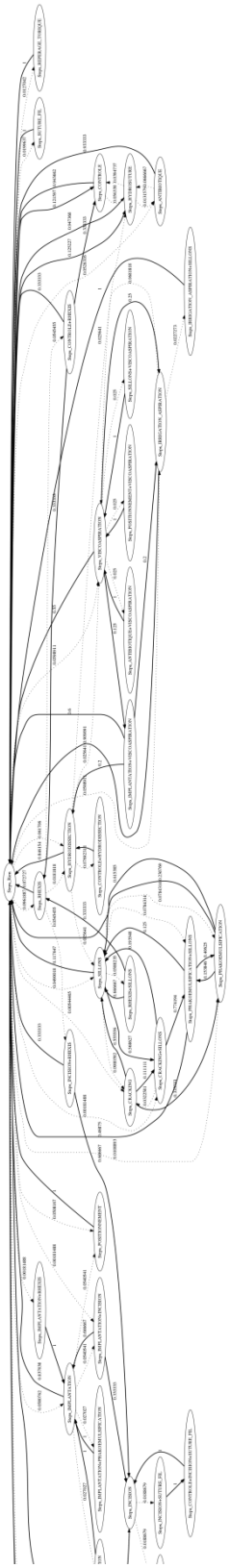
D. Martiano, K. Charrière, M. Lamard, and B. Cochener, “Indexing of cataract surgery video by content based video retrieval,” Acta Ophthalmologica, vol. 92, no. s253, pp. 0–0, 2014.

Annexes

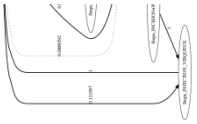
Annexes

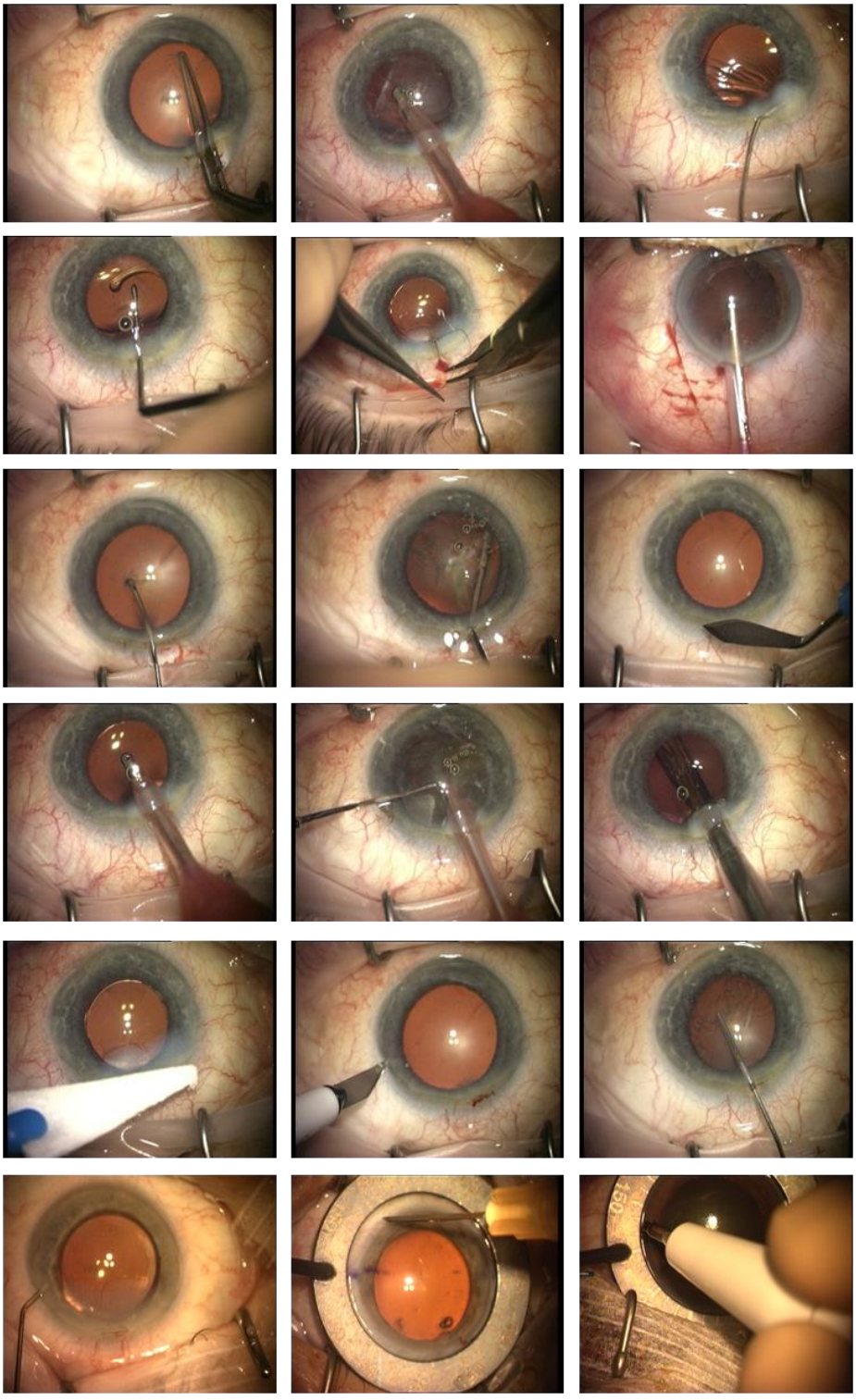


Annexe 1. Arbre construit selon la méthode d'apprentissage supervisée adaptée de la méthode de Piciarelli et al. (1), à partir de 41 vidéos



Annexe 2. Diagramme de transition des étapes chirurgicales





Annexe 3. Images des différents instruments qu'il est possible rencontrer dans nos vidéos de chirurgie de la cataracte

Résumé

L'objectif de cette thèse est de fournir aux chirurgiens des aides opératoires en temps réel. Nous nous appuyons pour cela sur des vidéos préalablement archivées et interprétées. Pour que cette aide soit pertinente, il est tout d'abord nécessaire de reconnaître, à chaque instant, le geste pratiqué par le chirurgien. Ce point est essentiel et fait l'objet de cette thèse.

Différentes méthodes ont été développées et évaluées, autour de la reconnaissance automatique du geste chirurgical. Nous nous sommes appuyés sur des méthodes de catégorisation (recherche des cas les plus proches basée sur l'extraction du contenu visuel) et des modèles statistiques du processus chirurgical. Les réflexions menées ont permis d'aboutir à une analyse automatique de la chirurgie à plusieurs niveaux de description. L'évaluation des méthodes a été effectuée sur une base de données de vidéos de chirurgies de la cataracte, collectées grâce à une forte collaboration avec le service d'ophtalmologie du CHRU de Brest.

Des résultats encourageants ont été obtenus pour la reconnaissance automatique du geste chirurgical. Le modèle statistique multi-échelles développé permet une analyse fine et complète de la chirurgie. L'approche proposée est très générale et devrait permettre d'alerter le chirurgien sur les déroulements opératoires à risques, et lui fournir des recommandations en temps réel sur des conduites à tenir reconnues. Les méthodes développées permettront également d'indexer automatiquement des vidéos chirurgicales archivées.

Mots-clés : Recherche de vidéos par le contenu, Modélisation multi-échelles, Analyse du processus chirurgical, Modèles de Markov, Champs markoviens conditionnels, Réseau bayésien

Abstract

Huge amounts of medical data are recorded every day. Those data could be very helpful for medical practice. The LaTIM has acquired solid know-how about the analysis of those data for decision support. In this PhD thesis, we propose to reuse annotated surgical videos previously recorded and stored in a dataset, for computer-aided surgery. To be able to provide relevant information, we first need to recognize which surgical gesture is being performed at each instant of the surgery, based on the monitoring video. This challenging task is the aim of this thesis.

We propose an automatic solution to analyze cataract surgeries, in real time, while the video is being recorded. A content based video retrieval (CBVR) method is used to categorize the monitoring video, in combination with a statistical model of the surgical process to bring contextual information. The system performs an on-line analysis of the surgical process at two levels of description for a complete and precise analysis. The methods developed during this thesis have been evaluated in a dataset of cataract surgery videos collected at Brest University Hospital.

Promising results were obtained for the automatic analysis of cataract surgeries and surgical gesture recognition. The statistical model allows an analysis which is both fine-tuned and comprehensive. The general approach proposed in this thesis could be easily used for computer aided surgery, by providing recommendations or video sequence examples. The method could also be used to annotate videos for indexing purposes.

Keywords : Content based video retrieval, Multilevel statistical model, Surgical process model, Markov models, Conditional random fields and Bayesian networks

TELECOM Bretagne

n° d'ordre : 2015telb0380

Télécom Bretagne

Technopôle Brest-Iroise - CS 83818 - 29238 Brest Cedex 3

Tél : + 33(0) 29 00 11 11 - Fax : + 33(0) 29 00 10 00